Detecting and Quantifying Causality from Time Series of Complex Systems

How information theory can help in discovering interaction mechanisms in the climate system

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m
(Dr. rer. nat.)
im Fach Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Dipl.-Phys. Jakob Gerhard Bernhard Runge

Präsident der der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. Prof. Dr. hc. Jürgen Kurths
2. Prof. Dr. Holger Kantz
3. PD Dr. Niels Wessel

Tag der mündlichen Prüfung: 5. August 2014

To my parents

Abstract

Today's scientific world produces a vastly growing and technology-driven abundance of time series data of such complex dynamical systems as the Earth's climate, the brain, or the global economy. In the climate system multiple processes (e.g., El Niño-Southern Oscillation (ENSO) or the Indian Monsoon) interact in a complex, intertwined way involving teleconnections and feedback loops. Understanding the *causal* interactions of such processes presents a major challenge for multidisciplinary scientific research, motivated not only by scientific curiosity, but by the urgent need to better understand anthropogenic climate change. The statistical analysis of measurements and observations to test or generate hypotheses on causal associations between dynamical processes constitutes an important goal in itself, but also offers a framework to validate and inspire physical models.

In this thesis, two main research questions are addressed: (i) How can general causal interactions be practically *detected* from multivariate time series? (ii) How can the *strength* of causal interactions between multiple processes be *quantified* in a well-interpretable way? In pursuing the first question, the causal notion aims at distinguishing direct from indirect interactions and common drivers and can be practically implemented in the framework of *conditional independence tests* extending *Granger causality* to general statistical associations. The second research question aims at a statistically and physically *well-interpretable* concept allowing to quantify different aspects of coupling mechanisms between two or more processes. Information theory is ideally suited to implement these goals and is the main framework harnessed in this thesis.

In the first part of this thesis, the theory of detecting and quantifying causal interactions is developed alongside with the important practical issues of estimation. The main challenge for the application of information-theoretic measures like *conditional mutual information* for high-dimensional multivariate data is the *curse of dimensionality*. The main contributions of this thesis to the first research question are (i) the combination of an advanced estimator of conditional mutual information with a causal inference algorithm that alleviates this curse. (ii) A proposed solution to overcome the important problem of autocorrelations being ubiquitous in time series data especially from climate. To address the second research question, this thesis contributes (iii) a physically motivated, information-theoretic formalism to quantify coupling mechanisms between two subprocesses, and also interactions between multiple subprocesses of a multivariate process, allowing to identify through which *causal paths* a complex interaction mechanism is mediated. The formalism is extensively tested numerically and substantiated by rigorous mathematical results.

In the second part of this thesis, the novel methods are applied to test and generate hypotheses on causal interactions in climate time series covering the 20th century up to the present. The results yield insights on an understanding of the Walker circulation and teleconnections of the ENSO system, for example with the Indian Monsoon. Further, in an exploratory way, a global surface pressure dataset is analyzed to identify key processes that drive and govern interactions in the global atmosphere. Finally, it is shown how quantifying interactions can be used to determine possible structural changes, termed *tipping points*, and as optimal predictors, here applied to the prediction of ENSO.

Zusammenfassung

Der technologische Fortschritt hat in jüngster Zeit zu einer großen Zahl von Zeitreihenmessdaten über komplexe dynamische Systeme wie das Klimasystem, das Gehirn oder das globale ökonomische System geführt. Beispielsweise treten im Klimasystem Prozesse wie *El Niño-Southern Oscillation* (ENSO) mit dem indischen Monsun auf komplexe Art und Weise durch Telekonnektionen und Rückkopplungen in Wechselwirkung miteinander. Das Verständnis der *kausalen* Wechselwirkungen dieser Prozesse stellt eine große Herausforderung für multidisziplinäre wissenschaftliche Forschung dar, motiviert nicht nur durch wissenschaftliche Neugier, sondern auch, um ein verbessertes Verständnis über den anthropogenen Klimawandel zu erlangen. Die statistische Analyse von Messdaten zum Testen und Aufstellen von Hypothesen über kausale Wechselwirkungen ist einerseits ein Ziel an sich, erlaubt aber darüber hinaus auch physikalische Modelle zu konstruieren und zu validieren.

Diese Dissertation verfolgt zwei Hauptfragen: (i) Wie können, ausgehend von multivariaten Zeitreihen, kausale Wechselwirkungen praktisch *detektiert* werden? (ii) Wie kann die *Stärke* kausaler Wechselwirkungen zwischen mehreren Prozessen in klar interpretierbarer Weise *quantifiziert* werden?

Zur Beantwortung der ersten Frage ist die Unterscheidung zwischen direkten und indirekten oder durch eine gemeinsame Ursache entstehenden Wechselwirkungen wichtig. Diese Frage der Kausalität lässt sich mit *bedingten Unabhängigkeitstests* umsetzen, welche das Konzept der *Granger-Kausalität* auf allgemeine statistische Abhängigkeiten erweitern. Die zweite Frage bezieht sich auf ein statistisch wie auch physikalisch *klar interpretierbares* Konzept, welches erlaubt, unterschiedliche Aspekte eines Kopplungsmechanismusses zwischen zwei, wie auch zwischen mehreren Prozessen statistisch möglichst parameterfrei zu quantifizieren. Zur Beantwortung dieser beiden Fragen beizutragen ist die Informationstheorie ideal geeignet und bildet die konzeptionelle Grundlage dieser Arbeit.

Im ersten Teil der Arbeit werden die Theorie zur Detektion und Quantifikation kausaler Wechselwirkungen (weiter-)entwickelt und wichtige Aspekte der Schätztheorie untersucht. Die größte Herausforderung für die Anwendung informationstheoretischer Maße, wie der *bedingten Transinformation*, ist der *"Fluch der Hochdimensionalität"*. Die wichtigsten Beiträge dieser Dissertation zur ersten Forschungsfrage sind (i) die Kombination von verbesserten Schätzern der bedingten Transinformation mit einem kausalen Algorithmus, der das Problem der Hochdimensionalität mindert. (ii) Eine Lösung des wichtigen Problems der Autokorrelation bei der Schätzung von kausalen Abhängigkeiten, welches häufig in Zeitreihen, insbesondere aus dem Klimabereich, vorkommt. Zur Beantwortung der zweiten Forschungsfrage wird (iii) ein physikalisch motivierter, informationstheoretischer Ansatz vorgeschlagen, mit dessen Hilfe Kopplungsmechanismen zwischen zwei, wie auch zwischen mehreren Teilprozessen eines multivariaten Prozesses quantifiziert werden können. In letzterem Fall erlaubt dies beispielsweise zu bestimmen, auf welchem *kausalen Pfad* ein physikalischer Mechanismus vermittelt wurde. Der Formalismus wird umfangreich numerisch untersucht und durch analytische Resultate untermauert.

Im zweiten Teil der Arbeit werden die entwickelten Methoden angewandt, um Hypothesen über kausale Wechselwirkungen in Klimadaten der vergangenen

hundert Jahre zu testen und zu generieren. Die Ergebnisse geben Aufschluss über die *Walker-Zirkulation* im Pazifik und Telekonnektionen von ENSO, beispielsweise mit dem indischen Monsun. In einem zweiten, eher explorativen Schritt wird ein globaler Luftdruck-Datensatz analysiert, um wichtige treibende Prozesse in der Atmosphäre zu identifizieren. Abschließend wird aufgezeigt, wie die Quantifizierung von Wechselwirkungen Aufschluss über mögliche qualitative Veränderungen in der Klimadynamik (*Kipppunkte*) geben kann und wie kausal treibende Prozesse zur optimalen Vorhersage von Zeitreihen genutzt werden können.

List of publications

Published papers

P1 **J. Runge**, V. Petoukhov, and J. Kurths, *Quantifying the strength and delay of climatic interactions: the ambiguities of cross correlation and a novel measure based on graphical models*, Journal of Climate 27(2), 720-739 (2014)

P2 **J. Runge**, J. Heitzig, N. Marwan, and J. Kurths, *Quantifying Causal Coupling Strength: A Lag-specific Measure For Multivariate Time Series Related To Transfer Entropy*, Physical Review E 86, 061121 (2012)

P3 **J. Runge**, J. Heitzig, V. Petoukhov, and J. Kurths, *Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy*, Physical Review Letters 108, 258701 (2012)

P4 C.F. Schleussner, **J. Runge**, J. Lehmann, and A. Levermann *The role of the North Atlantic overturning and deep-ocean for multi-decadal global-mean-temperature variability*, Earth System Dynamics 5, 103-115 (2014)

P5 B. Pompe and **J. Runge**, *Momentary Information Transfer as a Coupling Measure of Time Series*, Physical Review E 83, 051122 (2011)

P6 G. Balasis, R. V. Donner, S. M. Potirakis, **J. Runge**, C. Papadimitriou, I. A. Daglis, K. Eftaxias, and J. Kurths, *Statistical mechanics and information-theoretic perspectives on complexity in the Earth system*, Entropy special issue "Advances in Applied Statistical Mechanics" 15, 2844-4888 (2013)

P7 J. Hlinka, D. Hartman, M. Vejmelka, **J. Runge**, N. Marwan, J. Kurths, and M. Paluš, *Reliability of Inference of Directed Climate Networks Using Conditional Mutual Information*, Entropy 15(6), 2023-2045 (2013)

Preprint

R1 **J. Runge** *On the graph-theoretical interpretation of Pearson correlations in a multivariate process and a novel partial correlation measure*, arXiv:1310.5169v1 [math.ST].

Acknowledgements

I am deeply thankful to Prof. Jürgen Kurths for giving me the freedom to explore my ideas and for his support in manifold ways and the Potsdam Institute for Climate Impact Research (PIK) for providing a unique environment for scientific research on this beautiful Telegraphenberg.

I am indebted to the German National Academic Foundation (Studienstiftung des deutschen Volkes) for generously supporting my work financially and through inspiring summer schools and in manifold other ways for the past ten years. Additional funding was provided by the German Environmental Foundation (DBU), the DFG grant "KU34-1", and the DFG research group 1380 "HIMPAC".

I thank Bernd Pompe for introducing me to information theory which has shaped my interest since my Diploma thesis, Jobst Heitzig for numerous discussions, critical comments and pathological mathematical examples, Vladimir Petoukhov for sharing his immense knowledge on climate physics, and Barnabas Poczos from Carnegie Mellon University (Pittsburgh) for inspiring scientific exchange on estimation problems. For patiently listening to my ideas, productive interactions, discussions, and collaborations, I thank Alexander Radebach, Matthias Mengel, Jonathan Donges, Jakob Zscheischler, Lara Neureither, Norbert Marwan, Reik Donner, Anders Levermann, Carl Schleussner, Nishant Malik, my great office mates Lyuba Tupikina and Veronika Stolbova, Nora Molkenthin, and the whole Jürgen Kurths' group as well as other colleagues all over PIK.

I thank Roger Grzondziel, Ciaron Linstead, and Norbert Marwan for their help and support with using the IBM iDataPlex Cluster at PIK and for making available my software package *TiGraMITe*.

For their comments, suggestions, and proofreading efforts concerning various parts of this thesis, I thank Jürgen Kurths, Reik Donner, Jonathan Donges, Georg Runge, Jakob Zscheischler, Lara Neureither, Jobst Heitzig, Friederike Fröhlich, Alexander Radebach, Steffen Brunner, Andres Martinez, Norbert Marwan, Bernd Pompe, Manuela Runge, Kira Rehfeld, and Lyuba Tupikina.

Deep appreciation goes to my parents and my brother. Finally, thank you, R., for inspiring my life.

Contents

List of Figures

List of Tables

List of frequently used mathematical symbols

α	Significance level
$\mathcal{C}_{X_{t-\tau}\to Y_t}$	Set of intermediate nodes (including $X_{t-\tau}$) on causal paths from $X_{t-\tau}$ to Y_t in time series graph (Eq. (2.13))
$\langle\cdot\rangle$	Time or ensemble average
D_Z	Dimensionality of variable Z
η	Independently identically distributed noise
$\mathcal{G}$	Time series graph (see Sect. 2.4)
$\Gamma_{\mathbf{X}}$	Covariance matrix of multivariate process $\mathbf{X}$
h	prediction step ahead (see Chapter. 7)
$H(\,\cdot\,)$, $H(\,\cdot\,,\,\cdot\,)$, $H(\,\cdot\,\vert\,\cdot\,)$	Continuous marginal, joint and conditional Shannon entropy (see Sect. 3.2)
$I(\,\cdot\,;\,\cdot\,\vert\,\cdot\,)$	Conditional mutual information (if no '|' is given, the mutual information is meant, see Sect. 3.2)
$\mathcal{I}(\,\cdot\,;\,\cdot\,;\,\cdot\,\vert\,\cdot\,)$	Conditional interaction information (see Sect. 3.2)
k	Nearest-neighbor parameter of CMI estimator (see Sect. 4.2)
N	Number of processes
$\mathcal{N}_\cdot$, $\mathcal{N}(\cdot)$	Contemporaneous neighbors in time series graph (Eq. (2.11))
p	Order of model or number of predictors
$\mathcal{P}_\cdot$, $\mathcal{P}(\cdot)$	Parents of node(s) in time series graph (Eq. (2.10))
$\breve{\mathcal{P}}_{t+h}$	Predictors of process Y at time $t+h$ ahead (Eq. (7.4))
$\rho(\,\cdot\,,\,\cdot\,\vert\,\cdot\,)$	Partial correlation (if no '|' is given, the cross correlation is meant, see Sect. 3.3)
Σ, σ^2	Covariance matrix and variance of noise
t, τ, $\tau_{\max}$	Time, time lag, maximum time delay
$\mathcal{T}$, T	Set of time indices, length of this set
$\mathbf{X}_t$, $\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots)$	Multivariate process at time t and its past vector

List of abbreviations

AIR	All India Rainfall index (see Sect. 6.6)
ATL	Climate index in tropical Atlantic (see Sect. 6.3)
AUC	Area under receiver operating characteristic
CMI	Conditional mutual information (see Sect. 3.2.2)
CPAC	Climate index in tropical Central Pacific (see Sect. 6.3)
DMI	Indian Ocean Dipole mode
DTE	Decomposed transfer entropy (see Sect. 3.4.2)
EEUR	Climate index in eastern Europe (see Sect. 6.3)
ENSO	El Niño-Southern Oscillation
EPAC	Climate index in tropical East Pacific (see Sect. 6.3)
IIP	Interaction information (along paths), see Sect. 3.5.2
ISM	Indian Summer Monsoon
ITP	Information transfer along paths (see Sect. 3.5.1)
ITX	Information transfer from X (see Sect. 3.5.1)
ITY	Information transfer to Y (see Sect. 3.4.4)
LINK	Link-defining CMI (see Sect. 3.4.3)
MI	Mutual information (see Sect. 3.2.2)
MII	Momentary interaction information (see Sect. 3.5.2)
MIT	Momentary information transfer (see Sect. 3.4.5)
MITP	Momentary information transfer along paths (see Sect. 3.5.1)
NAO	North Atlantic Oscillation
Nino3	Climate index in tropical Pacific (see Sect. 6.3)
Nino3.4	Climate index in tropical Pacific (see Sect. 6.6)
PC algorithm	Peter Clark algorithm to infer time series graph (see Sect. 2.4.6)
PNA	Pacific/North American pattern
RMSE	Root-mean squared error
ROC	Receiver operating characteristic
SITY	State-Information transfer to Y (see Sect. 3.5.3)
SRMSE	Standardized root-mean squared error
SSA	Climate index in southern South America (see Sect. 6.3)
TE	Transfer entropy (see Sect. 3.4.2)
WEUR	Climate index in western Europe (see Sect. 6.3)
WPAC	Climate index in West Pacific (see Sect. 6.3)

Chapter 1.

Introduction

1.1. Main research questions

Let $\mathbf{X}$ be a complex system of which only measured multivariate time series are available. We want to know (Fig. 1.1):

1. How can we (practically) *detect* general causal interactions among the components of $\mathbf{X}$, including time lags?
2. How can we (practically) *quantify* the strength of causal interactions within $\mathbf{X}$ in a well-interpretable way?

The first question for causality is of much general interest in many fields of science underlying the search for the physical laws of nature and pursued through the iteration

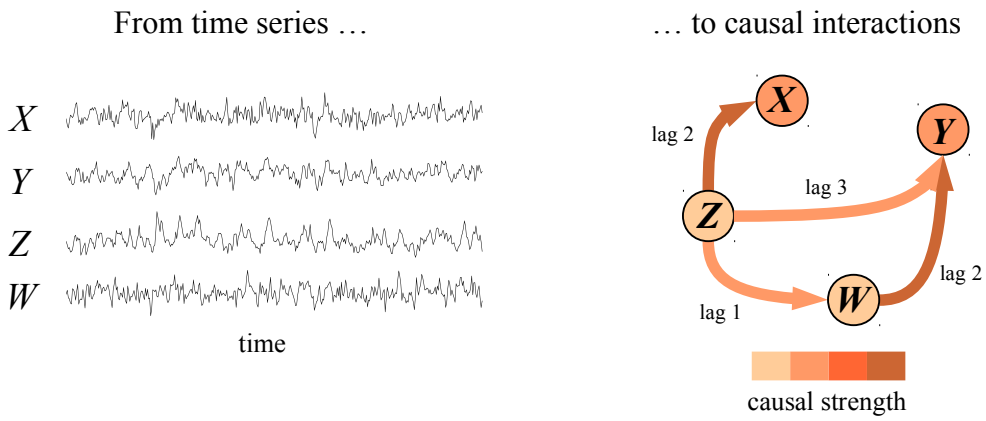


Figure 1.1.: Example of interactions between four processes, the research questions are: (1) Detecting causal links (arrows) including time lags (labels). For example, between X and Y there is no causal interaction even though they might be strongly correlated due to the common driver Z. (2) Quantifying the causal strength of links (arrow color) and more complex interactions like that between Z and Y which consists of a direct as well as an indirect coupling mechanism via W. The node color illustrates the internal causal strength of a process.

of theoretical hypotheses and laboratory experiments. But complex systems like the Earth system cannot be experimentally manipulated easily on a broad scale[1] and in other fields like neuroscience, ethical concerns limit this instrument of research. In Earth system science this has led to the development of computer models that simulate the climate system with the physical equations governing fluid motion and energy transfer. An alternative approach is to study a weaker form of causality in complex systems through the statistical analysis of measurements and observations.

For the statistical approach, one can view the subprocesses of a complex system $\mathbf{X}$ as nodes of a graph where the links denote inferred interactions. Often analyses, for example in climate and neuroscience (Tsonis et al., 2008; Donges et al., 2009a; Bullmore and Sporns, 2009), quantify interactions using pairwise statistical measures of association. Hans Reichenbach (1956) postulated that a statistical association between two processes X and Y implies that either there is a direct causal mechanism between the two or another process (or more) must act as a common driver Z (Fig. 1.1) which can be tested by measuring whether X and Y are statistically independent given Z. Similarly, an only indirect interaction inducing a correlation can give rise to misleading conclusions about the underlying mechanism. Therefore, a first step towards inferring *causal* interactions is to take into account other variables that might explain a statistical association. One of the first approaches relating statistical models to causality in this way is due to Wright (1921). The attribute '*general*' in our first question regarding causality refers to the demand that a statistical measure of association should pose as few assumptions as possible on an interaction and it should be sensitive to linear as well as nonlinear relationships (Rényi, 1959; Reshef et al., 2011). In dynamical systems such as the Earth system causal mechanisms on a macroscopic scale are not instantaneous, but interactions occur with *time delays* which constitute another important information about causal interactions in a complex system. Such an information also allows to improve statistical predictions. These demands of causality (including time delays) and generality formulate considerable challenges that have been addressed in different fields of science in the past decades as will be discussed in the next section. Finally, we demand such a method to be not only a theoretical concept, but also *practically* applicable.

In understanding the complex system Earth, the data-driven approach constitutes an important second pillar next to climate simulations (Von Storch and Zwiers, 2002). Causal inferences of climate time series can help to better understand the causes and effects of important subsystems such as the El Niño Southern Oscillation or the Indian Monsoon, whose impacts on agriculture and natural disasters are of paramount importance for billions of people (Philander, 1990; Pant and Kumar, 1997; Jin, 1997; Cane, 2005). In learning from past data, statistical analyses of causality can also help by improving simulation models needed for projections under global warming (Solomon, 2007). Further, on much shorter time scales, the knowledge of causal drivers

[1] Although humankind today exerts probably the largest uncontrolled experiment ever done in human history: climate change.

– even without an understanding of the mechanisms – could help to improve statistical predictions of these systems.

While the first question asks for a binary answer – causal or non-causal (possibly with an assessment of uncertainty) –, the second question already implies that there is more than one way of quantifying causal interactions and our aim is a statistically and physically *well-interpretable* concept. One aspect of the question of causal strength is the quantification of a link in the causal network of processes. This allows ranking the links and helps in deciding which interactions are important and should be included for example in a conceptual model, or which important interactions a complex model fails to reproduce. Further, one can view the influence of a subprocess X on Y in more detail as the combination of the *internal strength* of X, the *capacity* of the coupling mechanism, and the *susceptibility* of Y. There are many measures of the total influence between X and Y, but a separation of these different contributions in a well-interpretable way allows for a more comprehensive understanding of an interaction. For example, if a correlation between X and Y weakens over time, the reason can be that simply Y has become less susceptible while nothing has changed about the strength of X or the coupling mechanism. Figure 1.1 also depicts a more complex interaction mechanism between Z and Y which consists of a direct link as well as an indirect path via W. In this case, and also for purely indirect interactions and more connecting paths, one can ask how strong a coupling mechanism between Z and Y along all such paths is. This also includes the question *which* of the intermediate processes significantly mediates such a coupling.

Finally, also on a more global level, especially for larger complex systems, interesting questions arise. For example, to identify main driving and driven processes – sources and sinks of causal information transfer – and to quantify global properties such as the efficiency of information transfer (Latora and Marchiori, 2001). These questions have so far been addressed by analyzing the aforementioned functional networks (Bullmore and Sporns, 2009; Donges et al., 2009a) constructed from thresholded pairwise associations with the recently developed apparatus of statistical network theory (Newman, 2010). But many of these network measures such as the *average path length*, originating from the social sciences, are based on a different definition of links, i.e., two persons knowing each other as opposed to the statistical association between two processes. The question remains whether the inferences made from these 'static' network measures reflect the actual physics of causal information transfer. For example, if two processes are connected via several paths, the main coupling mechanism might not be the one via the shortest path. If measured over time, a precise quantification of causal information transfer could also turn out to be a good proxy of an *order parameter* that controls bifurcations of a complex system, termed *tipping points* in the climate system (Lenton et al., 2008). In sum, tools to quantify causal information transfer provide a more comprehensive understanding of the interaction structure and dynamics of complex systems.

1.2. Interactions between physics, statistics, and climate science

Different fields of science have studied complex systems from different perspectives. Making inferences from observational data is the primary goal of statistics and the more recent subfield of machine learning. Statisticians often view data as coming from some distribution while physicists take a more causal perspective using first principles and modeling natural processes using differential equations. Climate scientists, on the other hand, often want to understand mechanisms on a macro-level that parametrizes complex subprocesses. Physics has always inspired statistics (just like many other fields), for example, with the basic concept of entropy that flourished in *information theory* with the works of Claude Shannon (Shannon, 1948). Each perspective has its advantages and disadvantages and the combination of these different views has fertilized solutions for the research questions stated above.

Already the first question towards a causal interpretation constitutes a difficult problem involving fundamental philosophical aspects and immense statistical challenges. During the 20th century, statisticians have come up with a plethora of methods to model processes in different fields. From economically driven methods for predictions to the study of temperature variations in the tropics that led Sir Gilbert Walker to develop the theory of *autoregressive models* (Walker, 1923). In the 1960s, the physically inspired economist Clive Granger (Granger, 1969) gave a statistical definition of causality following an idea by Norbert Wiener. *Granger causality* is based on a statistical model for the prediction of a variable Y that is fitted to the data. If a variable X improves the prediction in this model, X is said to Granger-cause Y. Granger causality has been applied in many fields of science, especially economics (Granger, 1988), but in climate research the concept has only recently started to gain interest (Ebert-Uphoff and Deng, 2012b).

A statistical model always involves strong assumptions and the estimated absence of an interaction inferred with these model-based methods, therefore, does *not* imply that the processes are not interacting since only a certain class of causal mechanisms has been tested. In a more abstract model-free way, Pearl (2000) and Spirtes et al. (2000) have generalized Reichenbach's (Reichenbach, 1956) causal hypothesis between three processes mentioned above to the *Causal Markov Condition* between multiple processes which states that every process is *conditionally independent* of its *non-effects* given its *direct causes*. For example, in the causal chain $X \to Y \to Z$ the process Z is independent of its non-effect X given its direct cause Y. Information theory is ideally suited to implement the conditional independence approach which can be measured with quantities such as the *conditional mutual information*. But, if more than a few processes are to be tested as possible causal drivers, the actual estimation of conditional independence constitutes a challenge commonly faced in all high-dimensional inference problems and coined the *curse of dimensionality* by Bellman (1957). In the past decade, research in machine learning – made by philosophers (Spirtes and Glymour, 1991; Spirtes et al., 2000) – has come up with algorithms that

alleviate this curse and statistically motivated physicists have developed estimators of conditional mutual information (Frenzel and Pompe, 2007) that allow to better infer conditional independencies in a high-dimensional complex system. Time series constitute an especially difficult case because they are often strongly autocorrelated in time, which violates the very common assumption of independent samples making significance tests unreliable.

This thesis brings together and advances the above described recent developments from statistics, machine learning and physics in a multidisciplinary endeavor to address the two research questions on the detection and quantification of causal interactions stated above. As an approach to the first research question, the algorithm of causal inference (Spirtes and Glymour, 1991; Spirtes et al., 2000) is modified for the case of time series and combined with the recently developed estimators of conditional mutual information (Frenzel and Pompe, 2007) which enables the model-free inference of causal interactions also in a higher dimensional setting. This approach is analyzed from the underlying assumptions and limitations to estimation problems and the negative effect of autocorrelation on significance testing, which is largely overcome by the novel measures introduced in this thesis. Admitting to the inherent limitations of model-free techniques for common sample sizes in climate data, the framework is developed in parallel for linear measures – dropping the goal of generality.

For the second research goal to quantify causal interactions, a formalism based on a set of properties to quantify the causal strength of links as well as paths is presented. The formalism equally applies to linear and nonlinear interactions and suggests an intuitive and physically motivated interpretation that is substantiated with rigorous mathematical results. In particular, it will be shown how these measures can be used to disentangle the internal strength of X and the susceptibility of Y from the strength of their coupling mechanism as different contributors to an interaction. Further extending this approach, several measures are introduced that quantify the interaction between multiple processes and more global properties of information transfer in a complex system.

The framework can be used in two ways: As a confirmatory approach to test very specific hypotheses on the data implementing the statistical notion of conditional inference (Neyman and Scott, 1948; Reid, 1995; Amarasingham et al., 2012), and in an exploratory way to generate hypotheses based on the outcomes of an analysis. The improved significance tests help to gain confidence in the results obtained in this way.

The main focus of application for these novel methods is the complex system Earth. The results yield insights on an understanding of the Walker circulation and teleconnections of the El Niño-Southern Oscillation (ENSO) system, for example with the Indian Monsoon. Further, in an exploratory way, a global surface pressure dataset is analyzed to identify key processes that drive and govern interactions in the atmosphere. Finally, it is shown how causal interactions could be used to determine tipping points and as optimal predictors, here applied to the statistical prediction of ENSO.

Some parts of this thesis are the result of joint work published in the past years and some parts are first appearing in this thesis. In the rest of this thesis, I will mark at the beginning of chapters or sections whether they are at least partly based on published material and also mark in footnotes whether substantial contributions were made by co-authors and mostly stick to the first person plural or passive in the text.

1.3. Contents and arrangement of this thesis

The dissertation is divided into two parts introducing and discussing theory and estimation of causal interactions (Part I) and applications to climate and prediction (Part II).

Chapter 2, addressing the first research question, first provides an overview over the literature on inferring interactions from model-based approaches, via methods inspired from dynamical systems theory to the model-free information theory. Then, the approach taken in this thesis based on conditional independence is introduced and time series graphs are defined that encode the lag-specific causality of a multivariate process. Finally, theoretical algorithms of causal inference are introduced and the underlying assumptions and limitations are discussed. Chapter 3 is devoted to the second research question of quantifying causal interactions. After proposing properties of interaction measures, we review information theory as well as linear theory. Then we introduce and discuss common and novel measures suited to capture different aspects of causal interactions from links to paths and the interaction between multiple processes. Chapter 4 addresses the difficult problem of estimating information-theoretic functionals which are key to estimate conditional independence. Further, we analyze and improve tools for significance testing, especially for the case of autocorrelated time series, and discuss confidence bounds. Finally, we give examples and provide extensive numerical experiments on nonlinear models to validate our approach to estimate time series graphs. In Chapter 5 we move on to examples to develop an intuition for the interaction measures introduced in Chapter 3, culminating in coupling strength autonomy theorems that mathematically establish the basis for the physical interpretation which closes the chapter.

The second part covers Chapter 6 on climate interactions, where we study teleconnections from ENSO with new insights into its causal delays and demonstrate that our approach can reconstruct the major tropical mechanism of the Walker circulation. While the previous applications only considered few processes, we also study causal interactions in the network of major global sea-level pressure components constructed from the entire climatological field via a dimension reduction method. This allows to detect dominant drivers that affect many other processes from local to global scales. Further, we demonstrate how complicated pathways of coupling mechanisms can be disentangled. A final section on the ENSO – Monsoon interaction explores in more detail the case of non-stationary time series graphs and gives an outlook to determine tipping points of causal interactions. Chapter 7 demonstrates a second application, that of using the causal information to predict a time series, and demonstrates that

this approach can considerably improve statistical prediction methods. These insights are then applied to statistically forecasting ENSO.

Finally, Chapter 8 concludes this dissertation by recapitulating the main insights attained and sketching promising avenues for future research. Appendices provide analytical derivations, proofs, further applications and some more specialized topics.

Software In the course of this thesis, an open source software package written in the *Python* language (Van Rossum and Drake Jr, 1995) and *C* was developed that implements the novel methods. *TiGraMITe*, short for *time series graph and momentary information transfer estimation*[2], features a graphical user interface allowing also practitioners without programming knowledge to apply the framework to their data. Many of the figures in this thesis were produced with *TiGraMITe*.

[2] Available at http://tocsy.pik-potsdam.de/tigramite.php.

Part I.

Theory and Estimation

In this first part, the theory of detecting and quantifying causal interactions is developed alongside with the important practical issues of estimation. The novel measures are analytically and numerically studied on model systems and rigorous mathematical results are presented.

Chapter 2.

Determining causality from time series of complex systems

2.1. Introduction – from time series to interactions

Today large datasets of multivariate time series exist in fields such as the geosciences, ecology, neuroscience, physiology, genetics, and economics, representing such complex systems as the Earth, the brain, the human body, the genome and the global economy. These datasets often come from international organizations that collect, archive and make available data such as the National Oceanic and Atmospheric Administration (NOAA) in the United States for climate data or the Organization for Economic Co-operation and Development (OECD) for economic data. Main collaborative research projects such as the Human Brain Project in the European Union or the international Human Genome Project have fostered such data bases. These time series are measured in manifold ways from microarrays for gene expression data via the publication of economic data by countries or financial data in the stock market.

For spatially extended systems such as the brain and the Earth system, time series of various variables are measured using different techniques. For example, in electroencephalography, a dense net of electrodes attached to the scalp is used to measure electric activity. In climate science, observables such as surface air temperature and pressure have been globally measured using land stations and ships in the past and satellites in the recent decades. These scattered measurements are aggregated into gridded datasets using reanalysis methods (Kalnay et al., 1996).

Often the measured time series are not directly the entities of interest for inferring interactions, but aggregated variables are created using averaging, preprocessing or more complicated transformations that yield time series better representing a variable of scientific interest. For climate data (and also neuroscience), there are different approaches to use the multivariate time series measured at different locations for the inference of interactions. One approach is to directly view the individual time series as nodes of a network where the links are supposed to denote interactions. This approach has been developing rapidly and such networks have been termed functional brain networks (Bullmore and Sporns, 2009) and climate networks (Tsonis and Roebber, 2004; Donges et al., 2009a). These networks are typically derived by thresholding the matrix of associations to arrive at a binary adjacency matrix. Another approach is to employ a dimension reduction of such gridded datasets to first construct indices using

methods such as principal component analysis (called empirical orthogonal function analysis in the climate literature), singular spectrum analysis and more advanced techniques such as varimax rotated principal components (Wallace and Gutzler, 1981; Vautard and Ghil, 1989; Von Storch and Zwiers, 2002; Groth and Ghil, 2011). These components are then interpreted as representing different subprocesses of the climate system or the brain and associations between these indices can be measured. We will discuss the two methods and follow the latter approach in our climate analyses in Chapter 6. Both approaches relate a network to interactions by the mapping

$$\begin{aligned} \text{nodes} &\leftrightarrow \text{locations of time series or indices representing subprocesses}\,, \\ \text{edges} &\leftrightarrow \text{statistical associations}\,. \end{aligned}$$

Even though practitioners in these fields usually know the limits of invoked statistical methods towards a causal interpretation, they often are tempted to draw far-reaching conclusions from their analyses regarding coupling mechanisms. In Section 2.2 we review different statistical approaches from the climate literature in this respect. Section 2.3 reviews the literature on methods to measure causal interactions from model-based to model-free approaches, not all of which can cope with multivariate time series. Our approach, published with co-authors in Runge et al. (2012a); Runge et al. (2014), is introduced and discussed in Sect. 2.4 and the chapter closes with a discussion of further aspects of causal inference. This chapter covers the theoretical aspects of causality while the practical estimation of causal relations is studied in Chapter 4.

2.2. Measuring interactions

2.2.1. Zero-lag associations

The most basic approach to construct networks from multivariate time series is to estimate all pairwise Pearson correlations (introduced in Sect. 3.3, sensitive only to linear associations) or mutual informations (introduced in Sect. 3.2.2, sensitive also to nonlinear relationships) at lag zero, that is without shifting the time series against each other. Also measures like phase synchronization are frequently used (Pikovsky et al., 2003; Boccaletti et al., 2002; Arenas et al., 2008).

Donges (2012) gives an overview over climate network studies. Networks constructed from Pearson correlation and mutual information have been studied by Tsonis and Roebber (2004); Tsonis et al. (2006); Tsonis et al. (2008); Donges et al. (2009b); Donges et al. (2009a); Donges et al. (2011); Steinhaeuser et al. (2012) to name just a few in this rapidly evolving field. As mentioned in the introduction, often the resulting network topology is interpreted in terms of information transfer between different regions of the Earth. For example, Tsonis et al. (2008) relate network properties like *small worldness* (Newman, 2010) to the way in which information is transferred and

conclude on the stabilizing effect of "supernodes" associated with major dynamical patterns like El Niño-Southern Oscillation (ENSO) or the North Atlantic Oscillation (NAO).

But associations at lag zero cannot be interpreted in a directional way as the term "information transfer" implies. If some supernode is connected to many other regions, it could be that the other regions are actually driving the supernode or that their association is due to a common driving process, possibly simply a global warming trend. Note that for a directional interpretation it does not make a difference whether one uses linear or general nonlinear measures, or whether the link is based on significance or the value of the association measure (Paluš et al., 2011). An interpretation of directionality, not to speak of causality, necessitates at least that the measure involves a time-asymmetry.

2.2.2. Lagged associations

As a next step towards an assessment of directional links and to quantify the time lag of an association, lagged measures of association are invoked. In the construction of climate networks, this approach has been used, for example, in Yamasaki et al. (2008); Malik et al. (2012); Radebach et al. (2013).

Lagged correlation analysis has long been applied in climate research, popularized in the seminal works of Walker (1923); Walker (1924). It is used as a first step to gain insights into possible interaction mechanisms between different processes. Specifically, the cross correlation lag function is used to assess the time delay and to quantify the strength of the link mediated by a certain mechanism. To name just a few examples, Lanzante (1996) computed lag correlations of sea surface temperatures between different tropical regions to assess their mutual interaction. Klein et al. (1999) studied the mechanism by which the ENSO influences the Atlantic, the Indian Ocean and southern China. They inferred time delays between 3 to 6 months and suggest that changes in atmospheric circulation accompanying El Niño induce changes in cloud cover and evaporation which, in turn, increase the net heat flux entering these remote oceans. This was then postulated to be responsible for the surface warming. Gu and Adler (2011) investigated the impact of ENSO on tropical land surface temperatures and precipitation and find that the influence of ENSO on land precipitation has much shorter lags than the effect on land temperatures. They interpret this difference by suggesting: "This five-month time lag suggests a rough time scale needed for land surface air temperature to adjust because of the variations of surface energy budget caused by ENSO-associated circulation and precipitation anomalies." Hashizume et al. (2009) inferred a more complicated mechanism by investigating the impact of the Indian Ocean Dipole (DMI) on the Malaria risk in western Kenya. They find that "the 3- to 4-month lag in the positive association between DMI and the number of malaria cases coincided with the sum of the lag between DMI and rainfall (1 month) and the lag between rainfall and the incidence of malaria ($2-3$ months)." Also in Yamasaki et al. (2008); Gozolchiani et al. (2008), the lag and value of the cross correlation

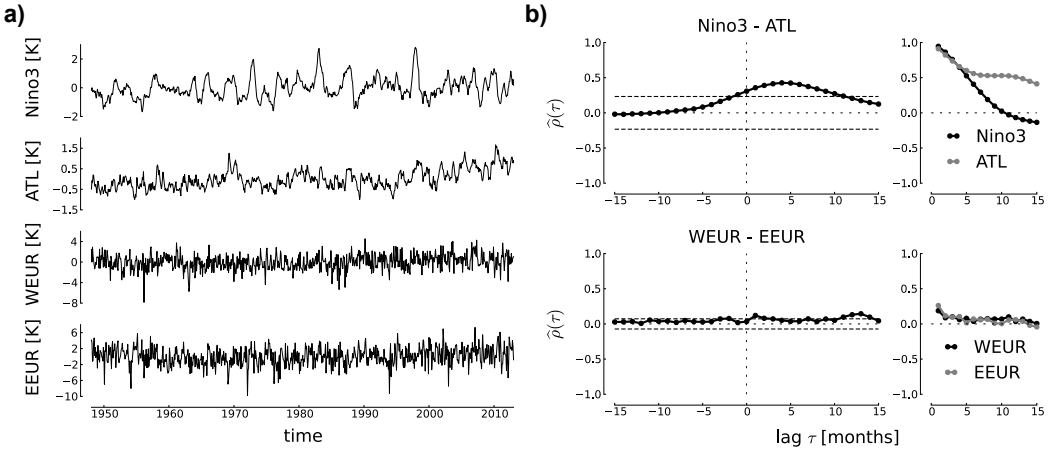


Figure 2.1.: (a) Time series and (b) estimated cross correlation (left) and autocorrelation (right) functions for monthly temperature anomalies from locations in the eastern tropical Pacific (Nino3), the tropical Atlantic (ATL), western (WEUR) and eastern Europe (EEUR), all regions are shown on the map in Fig. 6.3. Auto- and cross correlation are formally defined in Sect. 3.4.1. The lag one autocorrelation coefficients are 0.95 (Nino3), 0.91 (ATL), 0.19 (WEUR) and 0.26 (EEUR). In (b) the two-tailed $\alpha = 95\%$ significance threshold (dashed) for cross correlation is computed from uncorrelated Gaussian surrogate time series with the same autocorrelation coefficients and variances as the data. Note that for the autocorrelations the zero lag is not shown. The plots demonstrate apparent differences in the cross correlations that can be attributed to much stronger persistence in the tropical time series.

function at the maximum (divided by the standard deviation across a large range of lags) is interpreted as a measure of directionality and interaction strength.

These examples demonstrate that the delay at the maximum of the cross correlation function is used in interpreting the delay of the underlying physical mechanism that couples two processes. Also other lagged measures of association were proposed to determine lags in nonlinear processes, e.g., the mutual information (Granger and Lin, 1994). Apart from the analysis of time lags, the value of the cross correlation is commonly used as a measure of the effect of one process on another or a measure of the strength of a link or association, in line with the statistical interpretation of the square of correlation as the proportion of variance of one process that can be linearly represented by the other (Von Storch and Zwiers, 2002; Chatfield, 2013). These analyses are often accompanied by regressions.

But is it really justified to infer directionality and physical time lags from the maximum of the cross correlation function? How reliable is this method? The delay of what mechanism is actually measured? And how can the value of the cross correlation be interpreted *physically*?

To give a motivating example, we analyze cross correlation and autocorrelation

functions (formally defined in Sect. 3.4.1) for two very different pairs of monthly surface temperature *anomalies* for the period 1948–2012. In an anomaly time series the first moment of the seasonal cycle is removed prior to an analysis, details on the data are given in Chapter 6. In the first example in Fig. 2.1, Nino3 is the time series of the spatial average over the Nino3 region in the East Pacific and ATL is the average over a region in the tropical North Atlantic. In the second example, the cross correlation between two time series from Western (WEUR) and Eastern (EEUR) Europe is studied (all regions are shown on the map in Fig. 6.3). Figure 2.1(a) shows the time series and in (b) the cross correlation and autocorrelation functions are plotted. Several observations are apparent from Fig. 2.1(b): The peak of the tropical cross correlation with its maximum $\max_\tau \widehat{\rho}(\tau) = 0.43$ at lag $+4$ months is higher and broader than those of the European cross correlation which has values above the significance threshold only at lag $+1$ month ($\widehat{\rho} = 0.12$) and at around $+12$ to $+13$ months ($\widehat{\rho} = 0.13 - 0.14$). A correlation between the East Pacific and the tropical Atlantic is also reported in Lanzante (1996), where a lag of around 6 months with a correlation of 0.34 was found.

Interpreting cross correlation as a measure of the strength of a link mediated via a climatic mechanism, we have to ask: Do these results imply, that the Pacific – Atlantic link over a distance of about 4.500-8.500 km (depending on whether the region's corners or centers are used) is stronger than the link in Europe with a distance of only 2.000-3.000 km? Can one infer that the mechanism in the tropics is present at the whole range of lags of 0 to $+11$ months, since these lags are significantly correlated? And does the still significant value at lag -1 imply that it is a bidirectional interaction? Does the mechanism to transfer the Pacific anomalies take 4 months to reach the Atlantic? Can this be called a flow of information?

One explanation for the differences between the tropical and European correlations could be the much stronger persistence, that is, autocorrelation in the tropical time series as can be seen from the slowly decaying autocorrelation functions (Fig. 2.1(b)). Generally, often climatological time series exhibit these serial correlations or serial dependencies (Von Storch and Zwiers, 2002). Especially in tropical temperature time series, where the temperature at a given month very much depends on the temperatures of the previous months. Here the autoregressive lag one coefficients are 0.95 (Nino3) and 0.91 (ATL) in the tropics, but only 0.19 (WEUR) and 0.26 (EEUR) in the European midlatitudes. These differences seem to very much affect the cross correlation lag function, which raises the question how in particular a peak value and the lag at which the maximum occurs are to be interpreted. The influence of serial correlation on lagged correlation functions and regressions will be in detail analytically investigated in Section 5.2.1. It will be demonstrated how this influence can mislead conclusions about time delays and the direction of influence and how it also obscures a quantification of the interaction mechanism and, therefore, misguides a physical interpretation. In Sect. 6.3, the climatic examples studied here will be re-examined using the novel methods developed in this thesis.

2.3. Measuring causal interactions

2.3.1. Definition of Granger causality

The previous section has motivated some doubt on overinterpreting lagged associations, for example, the significant correlations at negative lags in Fig. 2.1(b) do not exclude the possibility of an influence also in the opposite direction. The question of the direction of influence between time series has been advanced considerably by the seminal works of Clive Granger (Granger, 1969). As recapitulated in Amblard and Michel (2012), he was inspired by an article by Wiener (1956) that was communicated to him by Denis Gabor. Interestingly, Norbert Wiener already quoted examples from climate and neuroscience for his predictive approach to causality between time series.

To define Granger causality, let X, Y be stationary stochastic processes and we denote by $\sigma^2(Y_t|U_t^-)$ the variance of the residual of predicting the time series Y using the information in the entire universe U accumulated from the infinite past until the present, denoted by $U_t^- = (U_{t-1}, \ldots, U_{t-\infty})$, and by $\sigma^2(Y_t|U_t^- \setminus X_t^-)$ the corresponding error variance if X is excluded from this information set denoted by $\setminus$. Assuming stationarity, one can drop the time indices.

Definition 2.1 (Granger causality)**.** *If* $\sigma^2(Y|U^-) < \sigma^2(Y|U^- \setminus X^-)$*, then we say that* X Granger-causes Y*.*

This definition implies that there is some unique information in X relevant for Y. As Granger already notes: "The one completely unreal aspect of the above definition[s] is the use of the series U_t representing *all* available information". (Granger, 1969). In practice U is replaced by a limited set of observed time series $\mathbf{X}$ and the above definition reads X *Granger-causes* Y *with respect to* $\mathbf{X}$. Granger did not specify what prediction method, i.e., linear or nonlinear, should be used to determine σ^2, but the use of the variance to quantify the closeness of prediction restricts this notion of causality to a *causality in mean* (Granger, 1969). In this thesis we will use a more general notion in the framework of information theory that takes into account the whole distribution of the residual, not just the variance (Sect. 2.4). Further, note that there are approaches to measure a stronger form of causality in a non-predictive sense. These approaches rely on the concept of *causal calculus* and *interventions* (Pearl, 2000; Pearl, 2009), which assume that the system can be experimentally manipulated as discussed in Sect. 2.5. We do not assume this for our scope of applications to climate data and will use the term "causal" in the weaker sense of Granger causality throughout this thesis. We make explicit the assumptions of causal inference in Section 2.4.7.

2.3.2. Model-based methods

Since the seminal article of 1969, Granger's and other's works (Granger, 1988; Geweke, 1984) have triggered a whole literature on causal inference methods which led to model-based causality tests in economics (Sims, 1972; Hiemstra and Jones, 1994; Hamilton, 1994; Rothman, 1999; Fan, 2003), neuroscience (Ding et al., 2006; Gourévitch et al.,

2006; Eichler, 2006), physiology (Riedl et al., 2010) and also some in the climate literature (Kaufmann and Stern, 1997; Triacca, 2005; Smirnov and Mokhov, 2009) for the bivariate case while Granger causality for the multivariate case was only introduced recently by Ebert-Uphoff and Deng (2012b); Ebert-Uphoff and Deng (2012a).

Typically, in these tests a certain model class is assumed, such as the class of generalized additive models (Hastie and Tibshirani, 1986),

$$\mathbf{X}_t^i = \sum_{j=1}^{N} \sum_{\tau=1}^{\tau_{\max}} f_{i,j,\tau}(\mathbf{X}_{t-\tau}^j) + \eta_t^i, \tag{2.1}$$

where f are functions belonging to some class such as polynomials, N is the number of time series used and $\tau_{\max}$ is the maximum lag up to which the model is fitted. To test whether the subprocess $\mathbf{X}^j$ Granger-causes $\mathbf{X}^i$, the model of $\mathbf{X}^i$ is fitted with and without including $\mathbf{X}^j$ and the reduction of the residual error η is quantified by some statistic such as an F-test (Brockwell and Davis, 2009).

To name just a few, this approach has been extended in various variants – often in the physics literature – from a frequency decomposition (Chen et al., 2006; Detto et al., 2012) to causality between multivariate variables (Barrett et al., 2010). The model assumptions have been relaxed using nonlinear extensions via radial basis functions (Ancona et al., 2004), Kernel methods (Marinazzo et al., 2008; Zhang et al., 2012), canonical correlations (Wu et al., 2011), partial directed coherence (Nawrath et al., 2010; Sommerlade et al., 2009) or in the Fokker-Planck framework (Prusseit and Lehnertz, 2008). Also the fitting procedures have been extended to iterative prediction schemes (Zhao et al., 2012) and the Gaussian assumption of the noise has been relaxed in Shimizu and Hoyer (2006); Hyvärinen et al. (2008). In statistics the general framework of inferring causal relations by formulating models is called *structural equation modeling* which has first been used by Sewall Wright (Wright, 1921) and has been formally defined by Judea Pearl (Pearl, 2000).

But, as mentioned in the introduction, a statistical model always involves strong assumptions and the estimated absence of an interaction inferred with these model-based methods, therefore, does *not* imply that the processes are not interacting since only a certain class of causal mechanisms has been tested. In general this problem is called *model misspecification*. Also, not only are true causal links missed, but model-based approaches can also lead to an increased number of false positives, i.e., non-causal links (Peters et al., 2013), for example, if an undetected nonlinear driver is responsible for a spurious interaction. In Peters et al. (2013) model-based fitting procedures are combined with an algorithm that tries to avoid such wrong conclusions. On the other hand, model-based approaches have several advantages as we will discuss in Sect. 2.5 and see in Chapter 4 on estimation.

2.3.3. Synchronization and recurrence-based methods

Another type of assumption is invoked in the methods based on phase synchronization or generalized synchronization (Rulkov et al., 1995; Rosenblum et al., 1996; Pikovsky et al., 2003; Boccaletti et al., 2002; Arenas et al., 2008) following the idea that oscillations in dynamical systems can be excited by other dynamical systems (Schelter et al., 2006; Smirnov and Bezruchko, 2009; Nolte et al., 2008). These methods necessitate that a phase can be extracted from the time series, e.g. by the Hilbert-transform (Rosenblum et al., 1996). To this end, the signal must "circulate" in phase space which is observed for real world time series from various fields of science such as from the cardiovascular system (Schäfer et al., 1998), but also climate (Maraun and Kurths, 2005). For predominantly stochastic systems, however, such an analysis is not possible. Similarly, also within the framework of recurrence analysis (Marwan et al., 2007), methods have been developed (Romano et al., 2007; Zou et al., 2012; Feldhoff et al., 2012), but so far only to assess directionality and not for the general multivariate setting.

2.3.4. Phase-space based methods

Also inspired by the theory of dynamical systems are the methods based on phase-space reconstruction (Casdagli et al., 1991; Gibson et al., 1992; Kantz and Schreiber, 2003) whereby each univariate time series is first converted to a time series of *state vectors*

$$\vec{X}_t = (X_t, X_{t-d}, \ldots, X_{t-(m-1)d}) \tag{2.2}$$

utilizing Takens' theorem (Takens, 1981), where m is the embedding dimension and d the embedding delay. Many methods are then based on making a prediction of the dynamics in the reconstructed phase space of one of the processes using local model fitting (linear or nonlinear) (Chen et al., 2004; Schiff et al., 1996; Faes et al., 2008) or local entropy measures (Faes et al., 2011). This framework is presented as an interesting complementary approach to Granger causality in Sugihara et al. (2012). Granger causality (and information theory) is primarily suited for stochastic systems, but not applicable in general to nonlinear dynamical systems where Takens' theorem applies, in particular to the class of non-separable systems. Underlying also here is the idea of reconstructing a state space and assessing causality by exploiting asymmetries in the membership to a common dynamical system. Sugihara et al. (2012) show non-separability for two coupled logistic equations, but the same idea can be demonstrated for the trivial deterministic equations

$$\begin{aligned} X(t) &= aY(t-1) \\ Y(t) &= bX(t-1). \end{aligned} \tag{2.3}$$

Here $Y(t)$ can be rewritten purely in terms of its own past alone as $Y(t) = baY(t-2)$. In the framework of Granger causality this implies that X does not improve

the prediction of Y such that no Granger causality exists. Granger discussed the deterministic case already in his seminal work (Granger, 1969) and concludes that his definition does not pertain to the deterministic case.

State-space methods provide an interesting alternative avenue to the stochastic based statistical approaches. But it is not clear whether it is applicable to the complex systems of interest in this thesis with more than the few variables and that probably do not exhibit some low dimensional attractor that can be reconstructed with embeddings. Also, consider the case if dynamical noise is added to at least X in Eq. (2.3),

$$\begin{aligned} & X(t) = aY(t-1) + \eta_t^X \\ \implies \quad & Y(t) = bX(t-1) = b(aY(t-2) + \eta_{t-1}^X), \end{aligned} \tag{2.4}$$

then non-separability vanishes and X Granger-causes Y again. That is, dynamical noise plays a crucial role for the determination of Granger causality and for real systems one might assume dynamical noise to always be present. Dynamical noise is also the basis of our approach to quantify interactions (Sect. 3.1.3).

2.3.5. Information-theoretic methods

Wiener's idea of prediction improvement can be phrased as a problem of inferring conditional independence. This allows for a more general definition of causality than in Def. 2.1, using the Causal Markov Condition (Pearl, 2000; Spirtes et al., 2000) mentioned in the introduction. This will be elaborated on in the next section. Conditional independence can very naturally be quantified in the framework of information theory (Cover and Thomas, 2006), a field that was also inspired by Wiener. Information theory treats the variables of interest as random processes and measures like mutual information quantify statistical dependencies based on the joint and marginal distributions of these variables. But, as formally defined in the next chapter, mutual information is a symmetric measure not allowing for directional inferences without further steps. The ideas to break this symmetry were already studied early on by Marko (1973) who considered the *directed information* in the context of data transmission over discrete memoryless channels with feedback. His ideas were further developed by Massey (1990); Kramer (1998); Amblard and Michel (2011).

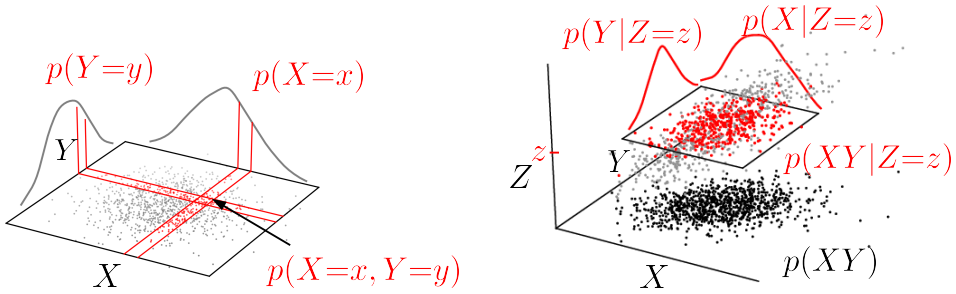
Transfer entropy (TE) (Schreiber, 2000; Paluš et al., 2001) has been widely used in different variants in the physics literature (Kaiser and Schreiber, 2002; Verdes, 2005; Hlaváčková-Schindler et al., 2007; Paluš and Vejmelka, 2007; Vejmelka and Paluš, 2008; Staniek and Lehnertz, 2008; Bahraminasab et al., 2008; Runge, 2010; Pompe and Runge, 2011; Hlinka et al., 2013; Faes et al., 2013; Kugiumtzis, 2013). Schreiber originally motivated transfer entropy as an alternative to lagged mutual information that takes into account shared information due to common history and input signals. TE for the direction $X \to Y$ is an information-theoretic distance measure between the transition probability that includes information from X and the one that excludes

it (the *entropy rate* or *source entropy* of the process as will be further elaborated on in Chapter 3). This already hints at a close connection with Granger causality (to which Schreiber actually made no reference) that can also be proven for the case of multivariate Gaussian processes (Barnett et al., 2009). Schreiber showed that TE is able to distinguish direct from indirect causality and, if external processes are taken into account, also common drivers without assuming any underlying model like the approaches discussed in the previous sections. TE will be in detail studied in Sections 3.4.2 and 5.2.2.

But despite these advantages, TE and similar approaches have mostly been applied in a bivariate setting as it is hard to estimate these measures reliably in high dimensions. In the definition of TE the Markov order of the process has to be known to estimate the entropy rate. In practice – to keep the dimensionality low – usually the first order is chosen, implying a somewhat arbitrary truncation. In this way TE does not appropriately account for longer delays typically occurring in real systems. Just like Granger causality, also TE is not designed to infer the important information of a coupling delay. An inference of the coupling delays based on a similar idea like TE has been done in Frenzel and Pompe (2007); Pompe and Runge (2011); Runge et al. (2012a); Wibral et al. (2013). In Frenzel and Pompe (2007), a first attempt to infer causality from multivariate times series is undertaken using a stepwise procedure to take into account the correct delays in multivariate applications.

In this thesis this iterative approach – optimized to keep the estimation dimension as low as possible – is fully formalized using ideas from machine learning and statistics together with the estimator by Frenzel and Pompe (2007). This method, therefore, is an attempt to overcome the problems that hindered a multivariate application of information-theoretic causality methods. To what extend this is feasible, is extensively discussed in this thesis. In the next section, we formally define our approach towards inferring causal interactions including the causal delays based on the idea of conditional independence. In this general framework, we can also embed parametric approaches with linear measures of conditional association (Sect. 3.3). Both approaches will be studied in this thesis.

For systems where the dynamical equations are known, Liang et al. (2005); Liang and Kleeman (2007b); Liang and Kleeman (2007a); Liang (2008); Majda and Harlim (2007); Liang (2013) have developed a rigorous formalism of information transfer between dynamical system components. Liang and Kleeman heuristically decompose the evolution of the marginal entropy into one part coming from the variable itself plus the information flow coming from another variable in the system. The first part is obtained using the system's equations with the other variable held fixed (using either the Frobenius-Perron operator for deterministic systems or the Fokker-Planck equation in the stochastic case) from which their notion of information flow follows. This approach yields an interesting microscopic view on information flow. However, our goal is to infer causality where the system is complex with the governing equations unknown. Nevertheless, to gain a better understanding of the novel introduced measures, we analytically study linear and nonlinear stochastic processes in Chapter 5.



a) Independence **b)** Conditional independence

Figure 2.2.: Scatter plots illustrating (conditional) independence. If in (a) the probability of the outcome (x, y) is equal to the product $p(x) \cdot p(y)$ for all (x, y), the processes are independent. Correspondingly, in (b) the same holds for the conditional probabilities given a value z.

2.4. Time series graphs

2.4.1. Conditional independence

Underlying the concept of a *time series graph* as formally defined in the next section is the theory of graphical models (Lauritzen, 1996). Graphical models visualize *conditional independencies* in a general multivariate process and can be used to draw inferences for Granger causal links (Bouezmarni et al., 2009). First, we briefly define and illustrate unconditional and conditional independence.

Two processes X and Y with joint probability density $p_{XY}(x, y)$ (in the following dropping the subscript) defined on $\mathbb{R}^2$ are *unconditionally independent* if and only if

$$p(x, y) = p(x) \cdot p(y) \quad \forall x,\, y \in \mathbb{R}^2\,, \tag{2.5}$$

where $p(x)$, $p(y)$ are the marginal densities. This is illustrated for a realization of a two-dimensional Gaussian process in Fig. 2.2(a).

Now consider three processes where X drives Z and Z drives Y as visualized in Fig. 2.3(a). Here X and Y are not directly, but indirectly interacting and in a bivariate analysis X and Y would be found to be dependent – implying that their correlation would be nonzero in the case of a linear dependency. The same holds for a common driver scheme in Fig. 2.3(b). If, however, the variable Z is included into the analysis, one finds that X and Y are independent *conditional* on Z, written as

$$X \quad \perp\!\!\!\perp \quad Y \quad | \quad Z. \tag{2.6}$$

This case is shown in a 3-D scatter plot in Fig. 2.2(b), where now the *conditional joint density* $p(x, y|z) \equiv \frac{p(x,y,z)}{p(z)}$ for every $z \in \mathbb{R}$ factorizes:

$$p(x, y|z) = p(x|z) \cdot p(y|z) \quad \forall x, y, z \in \mathbb{R}^3. \tag{2.7}$$

This concept will now be used to define (Granger-) causal time series graphs. Conditional independence can be estimated using information-theoretic functionals as discussed in Sect. 3.2. Typically Z will be a multivariate process, denoted by $\mathbf{Z}$, and the estimation of conditional dependence from finite time series presents a challenging problem. In Sections 4.2 and 4.3 we study in detail numerically how well conditional independence can be practically estimated for high-dimensional $\mathbf{Z}$ and with the further difficulty of *autocorrelated* time series typically occurring in climate data.

2.4.2. Definition of links

Time series graphs are based on the concept of conditional independence like graphical models and were introduced for the linear case by Dahlhaus (2000); Dahlhaus and Eichler (2003); Eichler (2005) and in a certain nonlinear generalization to phase synchronization by Schelter et al. (2006). Here, we introduce them for the general stochastic case via conditional independence, which has been termed *strong Granger causality* in Florens and Mouchart (1982); Eichler (2012).

Consider a multivariate process $\mathbf{X}$ of dimension N with set of components V. Then we define the *time series graph* $\mathcal{G} = (V \times \mathbb{Z}, E)$ of $\mathbf{X}$ as follows. As depicted in Fig. 2.4(a), the set of nodes in that graph consists of the set of components V at each time $t \in \mathbb{Z}$. That is, the graph is actually infinite. Compared to the general concept of graphical models (Lauritzen, 1996) for data without time-ordering, for time series graphs the time-dependence is explicitly used to define directional links in E. For convenience, we treat $\mathbf{X}$, $\mathbf{X}_t$, and $\mathbf{X}_t^-$ as sets of random variables here and use the difference symbol "\" for sets.

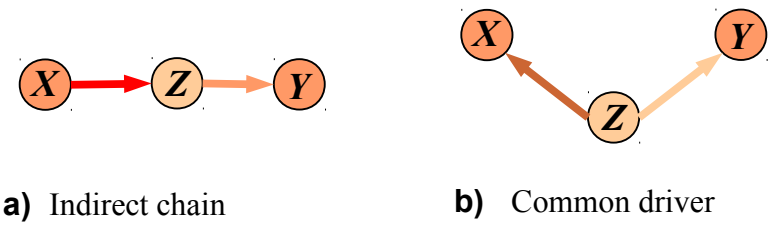


Figure 2.3.: Causality between three processes: (a) Indirect chain and (b) common driver system. The color of links underlines the difference between the graphical models approach which only assesses the existence of causal links and our approach to additionally quantify their strength as discussed in Chapter 3.

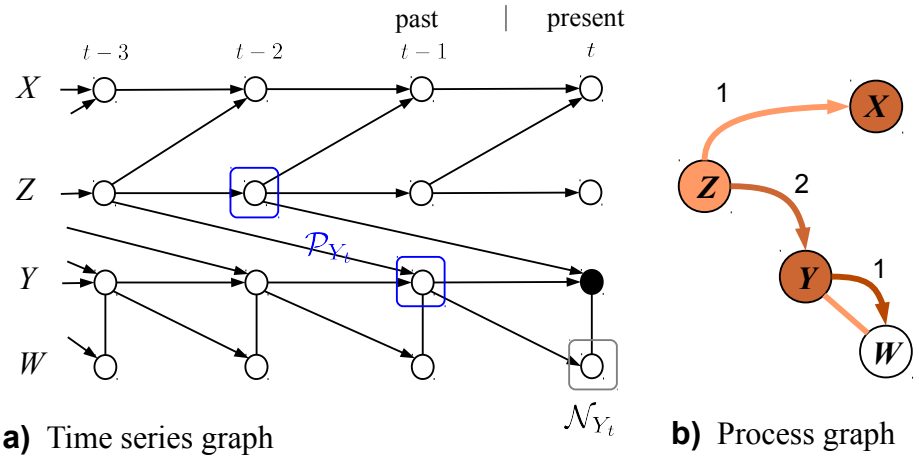


Figure 2.4.: (a) Time series graph. Each node corresponds to a lagged subprocess and due to stationarity, links for t imply links for all $t-1, t-2, \ldots$. Process Y_t (black node) has two parents (blue boxes, connected via incoming links from the past) and one neighbor W_t (grey box, connected with an undirected contemporaneous link) as defined in Eq. (2.10) and (2.11). (b) Process graph, which aggregates the information in the time series graph for better visualization. Labels denote the (possibly multiple) lags, link and node colors encode the coupling strength of some interaction measure and the autodependency strength, respectively, as defined in Chapter 3. Here W does not have any autodependency and the node color is white. Note, however, that unconditionally, there is a spurious dependency between W_t and its own past due to Y.

Definition 2.2. *Nodes $X_{t-\tau} \in \mathcal{G}$ and $Y_t \in \mathcal{G}$ are connected by a lag-specific* directed link *"$X_{t-\tau} \to Y_t$" pointing forward in time if and only if $\tau > 0$ and*

$$X_{t-\tau} \not\perp\!\!\!\perp Y_t \mid \mathbf{X}_t^- \setminus \{X_{t-\tau}\}, \tag{2.8}$$

i.e., if they are not *independent conditionally on the past of the whole process denoted by $\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots)$, which implies a lag-specific conditional dependence with respect to $\mathbf{X}$.*

If $Y \neq X$, the link "$X_{t-\tau} \to Y_t$" represents a *coupling at lag* τ, while for $Y = X$ it represents an *autodependency at lag* τ. Note that stationarity implies that "$X_{t-\tau} \to Y_t$" whenever "$X_{t'-\tau} \to Y_{t'}$" for any t' and analogously for contemporaneous links.

In Eichler (2012), the definition of links was given between components at *all* lags which does not allow for a lag-specific interpretation. While this might sometimes also not be desired, we include it because it allows for a more precise physical interpretation. Another difference to the literature on graphical models and the special case of *directed acyclic graphs* (DAGs) demanded for causal inferences (Spirtes and Glymour, 1991; Spirtes et al., 2000) is that we actually consider a *mixed graph* with

two different kinds of edges, directed and contemporaneous ones (Eichler, 2012). The definition of contemporaneous links is inspired from climate applications, where often an interaction measured from monthly time series is not lagged, but contemporaneous. We do not try to infer a causal relation in this case (which has been done in a model-based framework, e.g., in Chu and Glymour (2008); Peters et al. (2013)), but we believe that this information is better kept in the graph rather than left out.

Definition 2.3. *Nodes $X_t \in \mathcal{G}$ and $Y_t \in \mathcal{G}$ are connected by an undirected* contemporaneous link *"$X_t - Y_t$" if and only if*

$$X_t \not\perp\!\!\!\perp Y_t \;|\; \mathbf{X}^-_{t+1} \backslash \{X_t, Y_t\}, \tag{2.9}$$

where also the contemporaneous present $\mathbf{X}_t \backslash \{X_t, Y_t\}$ is included in the condition.

Sometimes such an undirected link might actually be causal, but the time sampling does not allow to distinguish cause and effect. This time sampling problem was already discussed in Granger (1969) in economics. For DAGs coming from data without a time order, the problem that several causal graphs are *Markov equivalent* with each other exists (Spirtes et al., 2000). For example, $X \to Y \to Z$ is Markov equivalent to $X \leftarrow Y \leftarrow Z$. In our case the time order determines the direction of the arrows.

2.4.3. Causal Markov property

Before we discuss how this graph can be estimated from time series, we give some theoretical relations between the graph $\mathcal{G}$ and the underlying process $\mathbf{X}$, most importantly the *Causal Markov Condition* (Spirtes et al., 2000) which provides a causal interpretation of time series graphs. The definition of this condition and the interaction measures in the next chapter is based on the important notion of the *parents* $\mathcal{P}_{Y_t}$ and the *neighbors* $\mathcal{N}_{Y_t}$ of a process Y_t in the time series graph. They are defined as

$$\mathcal{P}_{Y_t} \equiv \{X_{t-\tau} : \; X \in \mathbf{X}, \; \tau > 0, \; X_{t-\tau} \to Y_t\}, \tag{2.10}$$
$$\mathcal{N}_{Y_t} \equiv \{X_t : X \in \mathbf{X}, X_t - Y_t\}. \tag{2.11}$$

Note that also the past lags of Y can be part of the parents of Y_t. For example, in Fig. 2.4(a) the parents of Y_t are $\{Y_{t-1}, Z_{t-2}\}$ and the only neighbor is W_t. The parents of all subprocesses in $\mathbf{X}$ together with the contemporaneous links comprise the time series graph. In Chapter 7, we study in how far the parents can be used as optimal predictors for time series forecasting.

The Causal Markov Condition (Spirtes et al., 2000) now states that all $Y_t \in \mathbf{X}_t$ are independent of $\mathbf{X}^-_t \setminus \mathcal{P}_{Y_t}$ given the direct causes $\mathcal{P}_{Y_t}$. Spirtes et al. (2000) consider two more axioms connecting probabilities with causal graphs, namely the *Causal Minimality Condition* and *Faithfulness*. The former implies that no link in the graph can be removed without leading to a graph that violates the Causal Markov Condition. Faithfulness will be discussed in Sect. 2.4.7. The Causal Markov Condition can also

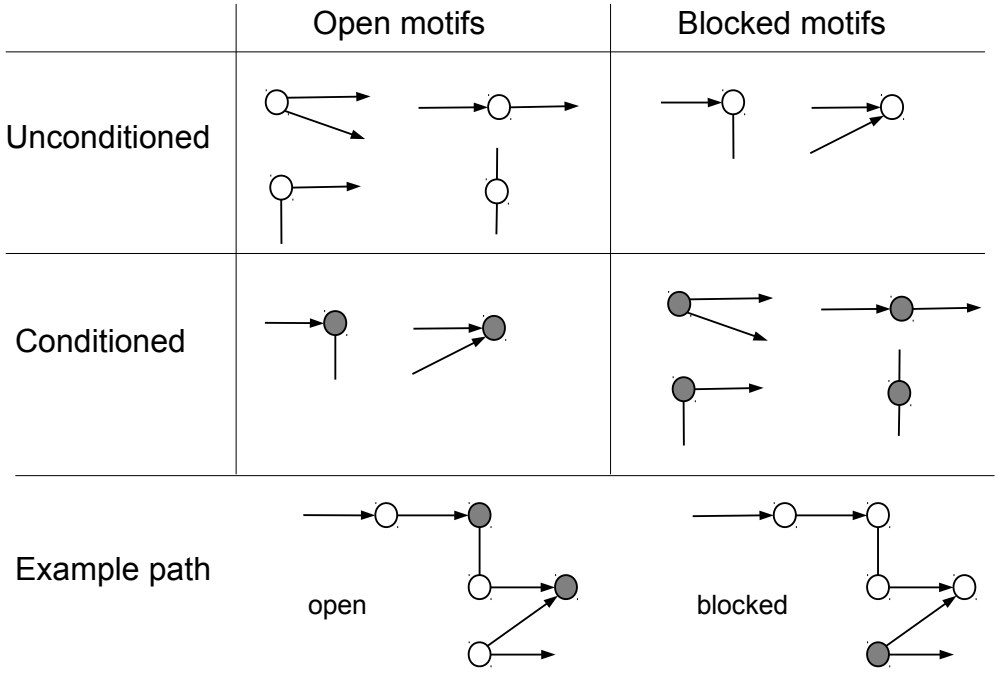


Figure 2.5.: Open and blocked motifs on a path. The six motifs in the upper two rows show all possibilities of motifs with two links (directed and contemporaneous). Nodes included in the conditioning set S in Eq. (2.12) are in grey. The bottom row shows an example of an open (left) and a blocked (right) path. One blocked motif is sufficient to block a whole path. Note that for a different definition of contemporaneous links discussed in Eichler (2012) (without including $\mathbf{X}_{t+1}$ in the condition), the unconditional motif $-o-$ becomes blocked and the conditional one becomes open.

be generalized to obtain independence relations not only for Y_t an its parents, but for general sets of nodes in the graph. To this end, an important notion is that of a *path*. Following Eichler (2012), a path π between two nodes $a, b \in V \times \mathbb{Z}$ is a sequence of edges $\pi = (e_1, \ldots, e_l)$ such that e_i is an edge between v_{i-1} and v_i for a sequence of vertices $v_0 = a, \ldots, v_l = b$. Each intermediate node with its adjacent edges forms a *motif*. Now we define *unconditioned open* and *unconditioned blocked* motifs as shown in Fig. 2.5. In the graphical models literature the intermediate nodes in open motifs are called *non-colliders* and in blocked motifs *colliders* (Eichler, 2012). We will see that unconditional open motifs can be related to unconditional dependence. Before, we also introduce the conditional extensions, just like conditional independence is an extension of independence. Let S be a subset of nodes from $V \times \mathbb{Z}$ that later forms the conditioning set in measures of conditional dependence. Then we define *unconditioned open* and *unconditioned blocked* motifs as shown in Fig. 2.5, i.e., the openness and blockedness is reversed if the intermediate node is conditioned on. Now

a path consisting only of (conditioned or unconditioned) open motifs is an *open path given* S and a path with at least one (conditioned or unconditioned) blocked motif is a *blocked path given* S. With these definitions we can now define *separation* in a graph following Eichler (2012) where this notion of separation is called *p-separation*[3].

Definition 2.4. *Two vertices a and b in a mixed graph $\mathcal{G}$ are* separated given a set S *if all paths between a and b are blocked given S. Similarly, two sets A and B in $\mathcal{G}$ are said to be separated given S if, for every pair $a \in A$ and $b \in B$, a and b are* separated given S. *This will be denoted by $A \bowtie B|S$.*

Then the Markov property states that separation in the graph yields conditional independence relations of the underlying process, i.e.,

$$A \bowtie B|S \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B|X_S. \tag{2.12}$$

This relation entails the Causal Markov Condition stated above because the set of parents $\mathcal{P}_{Y_t}$ (or any subset of $\mathbf{X}_t^-$ that contains $\mathcal{P}_{Y_t}$) separates Y_t from $\mathbf{X}_t^- \backslash \mathcal{P}_{Y_t}$ in the graph. This can easily be seen, because all parents are coming from nodes with directed links towards Y_t and since all of these are conditioned on, all paths are blocked. The algorithm used to estimate the time series graph (Sect. 2.4.6) makes use of this Markov property.

Further, we define a *directed path* which consists only of directed motifs $\rightarrow v_i \rightarrow$. We call these paths also *causal paths* and denote the processes along all directed paths including $X_{t-\tau}$ and excluding Y_t by

$$\mathcal{C}_{X_{t-\tau} \rightarrow Y_t} \equiv \{Z_{t-\tau_Z} : \; Z \in \mathbf{X}, \; \tau_Z > 0, \; X_{t-\tau} \rightarrow \ldots \rightarrow Z_{t-\tau_Z} \rightarrow \ldots \rightarrow Y_t\} \cup \{X_{t-\tau}\}, \tag{2.13}$$

where $\rightarrow \ldots \rightarrow$ denotes a causal path or a causal link.

With the Markov property Eq. (2.12) we can now understand why, for example, the cross correlation function shown in Fig. 2.1(b) has many "significant" values. Consider the time series graph shown in Fig. 2.4(a). There, the two nodes X_{t-1} and Y_t are connected by many open paths via the past (e.g., $X_{t-1} \leftarrow Z_{t-2} \rightarrow Y_t$, but also $X_{t-1} \leftarrow X_{t-2} \leftarrow Z_{t-3} \rightarrow Y_{t-1} \rightarrow Y_t$), making them unconditionally dependent. But, if we use the parents of Y_t as a conditional set, i.e., $S = \mathcal{P}_{Y_t}$, we see that all paths are actually blocked. This highlights an important aspect of causal lags or delays which are crucial to determine causal interactions in time series. For example, if we omit the autodependency Y_{t-1} from the conditional set S, then X_{t-1} and Y_t are *not* independent given only Z_{t-2}, because many paths are still open via Y_{t-1}. These subtle interactions can be captured with time series graphs that take into account autodependencies and delayed interactions.

[3]For directed acyclic graphs the notion of *d-separation* holds (Lauritzen, 1996).

2.4.4. Linear case – autoregressive models

While the definition of time series graphs was given for the large class of processes having the Markov property (Spirtes et al., 2000; Pearl, 2000), in this section we consider the case of a stationary N-variate discrete-time process defined as

$$\mathbf{X}_t = \sum_{s=1}^{p} \Phi(s)\mathbf{X}_{t-s} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \tag{2.14}$$

i.e., a vector autoregressive process (VAR) of order p where $\Phi(s)$ are $N \times N$ matrices of coefficients for each lag s and the N-vector ε is an independently identically distributed Gaussian random variable with zero mean and covariance matrix Σ. ε is sometimes referred to as the *innovation term*. Its variances on the main diagonal of Σ we denote by σ_i^2 and the covariances by σ_{ij} for $i \neq j$. This linear process will serve as an analytically solvable example for many properties of interaction measures discussed in this thesis. The model can be understood as a discrete-time sampling of a multivariate Ornstein-Uhlenbeck process studied in many branches of physical sciences (see Sect. 5.5.3).

For this model class, the directed and contemporaneous links of the corresponding time series graph are defined by non-zero entries in the coefficient matrix Φ and the inverse of the innovation covariance matrix Σ (Eichler, 2012):

$$X_{t-\tau} \to Y_t \quad \Leftrightarrow \quad \Phi_{YX}(\tau) \neq 0 \tag{2.15}$$

$$X_t - Y_t \quad \Leftrightarrow \quad (\Sigma^{-1})_{YX} \neq 0. \tag{2.16}$$

An alternative definition of contemporaneous links is based on non-zero entries in Σ_{YX} (Eichler, 2012).

As an example, consider the bivariate autoregressive model of order 1,

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \underbrace{\begin{pmatrix} a & 0 \\ c & b \end{pmatrix}}_{\Phi(1)} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{X,t} \\ \varepsilon_{Y,t} \end{pmatrix} \tag{2.17}$$

and $\Phi(s) = 0$ for $s > 1$. In Fig. 2.6 the corresponding time series graph is visualized.

2.4.5. Time series graphs for non-stationary processes

The parents and neighbors in time series graphs can also be defined for subsets of the time axis indices $\mathcal{T}$ to extend the concept of a time series graph to the non-stationary case. Denoting the subset of selected indices as $\mathcal{T}_Y \subseteq \mathcal{T}$, the definitions given in Eq. (2.8) then read

$$X_{t-\tau} \not\perp\!\!\!\perp Y_t \;\mid\; \mathbf{X}_t^- \setminus \{X_{t-\tau}\} \quad \forall t \in \mathcal{T}_Y, \tag{2.18}$$

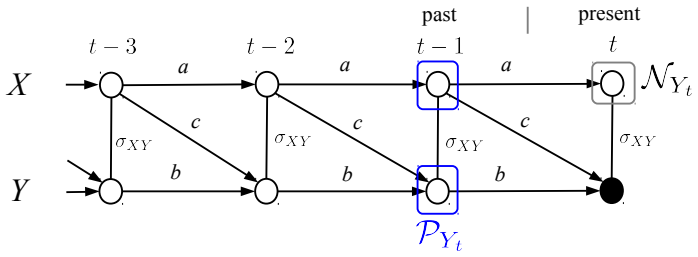


Figure 2.6.: Visualization of model Eq. (2.17) as a time series graph. The labels indicate the coefficients in the matrices $\Phi(1)$ and Σ. Note that a non-zero coefficient only determines the existence or absence of a link, but not a weight. Further, a non-zero σ_{XY} only defines a contemporaneous link in the bivariate case, while it is non-zero entries in $(\Sigma^{-1})_{YX}$ in the multivariate case. Due to stationarity, links for t imply links for all $t-1$, $t-2$, ... Process Y_t (black node) has one neighbor X_t (grey box) and two parents (blue boxes).

and correspondingly for contemporaneous links in Eq. (2.9). Non-stationarities of causal relations are very frequent in climate. For example, rainfall in India has different causal interactions with external processes during the Indian Summer Monsoon (ISM) from June to September than in the rest of the year (Pant and Kumar, 1997). In the midlatitudes seasonality is also important and often an interaction mechanism holds only for winter months and might even be reversed in other seasons. In the climate applications in Chapter. 6, the seasonality will be taken into account.

2.4.6. Causal algorithm

The estimation of graphical models is very similar to the problem of inferring the *directed acyclic graph* (Spirtes et al., 2000) of a set of random variables. To this end, the idea of the PC algorithm (named after its inventors Peter and Clark) is to iteratively unveil the links by testing for conditional independence between all possible pairs of nodes conditioned on iteratively more conditions and testing all combinations among them (Spirtes and Glymour, 1991; Spirtes et al., 2000). Thereby, the dimension stays as low as possible in every iteration step. We will discuss in Chapter 4 that higher dimensions give rise to a curse of dimensionality strongly affecting the reliability of inferring conditional independence. Since the PC algorithm was originally introduced to estimate graphical models where no information about time-order is assumed in the data, it consists of two steps: In the first step only undirected links are inferred, which are tested for directionality in the second step. But in our case of time series, the time ordering of nodes already provides the directionality and we omit the second step. Instead we estimate the contemporaneous links defined in Eq. (2.9) without trying to assess a directionality, which also circumvents the problem of Markov equivalence as discussed above. Further, we propose some modifications to speed up the performance discussed in Section 4.4.

The algorithm starts with no a priori knowledge about the links and iteratively learns the set of parents and neighbors for each Y. The union of parents together with the contemporaneous links then comprises the graph. Here we phrase the algorithm with some measure $I(X;Y|\mathbf{Z})$ able to estimate conditional independence $X \perp\!\!\!\perp Y|\mathbf{Z}$ for possibly multivariate $\mathbf{Z}$, which can be either information-theoretic estimators introduced in Chapter 4 or linear estimators like partial correlation.

For every Y, first we estimate unconditional dependencies $I(X_{t-\tau};Y_t)$ and initialize the preliminary parents $\tilde{\mathcal{P}}_{Y_t} = \{X_{t-\tau} : X \in \mathbf{X}, 0 < \tau \leq \tau_{\max}, I(X_{t-\tau};Y_t) > 0\}$. This set contains also indirect links which are now iteratively removed by testing whether the dependence between Y_t and each $X_{t-\tau} \in \tilde{\mathcal{P}}_{Y_t}$ conditioned on the incrementally increased set of conditions $\tilde{\mathcal{P}}^{n,i}_{Y_t} \subseteq \tilde{\mathcal{P}}_{Y_t}$ vanishes:

n. Iterate n over increasing number of conditions, starting with some $n_0 > 0$:

  n.i Iterate i through all combinations of picking n nodes from $\tilde{\mathcal{P}}_{Y_t}$ to define the conditions $\tilde{\mathcal{P}}^{n,i}_{Y_t}$ in this step, and estimate $I(X_{t-\tau};Y_t \,|\, \tilde{\mathcal{P}}^{n,i}_{Y_t})$ for all $X_{t-\tau} \in \tilde{\mathcal{P}}_{Y_t}$. After each step the nodes $X_{t-\tau}$ with $I(X_{t-\tau};Y_t \,|\, \tilde{\mathcal{P}}^{n,i}_{Y_t}) = 0$ are removed from $\tilde{\mathcal{P}}_{Y_t}$ and the i-iteration stops if all possible combinations have been tested.

  If the cardinality $|\tilde{\mathcal{P}}_{Y_t}| \leq n$, the algorithm converges, else, increase n by one and iterate again.

Once the parents of each process are known, the same algorithm for $\tau = 0$ can be used to infer the contemporaneous neighbors $\mathcal{N}_{Y_t} = \{X_t : X \in \mathbf{X}_t, X_t - Y_t\}$, where now undirected links are removed if $I(X_t;Y_t \,|\, \mathcal{P}_{Y_t}, \tilde{\mathcal{N}}^{n,i}_{Y_t}, \mathcal{P}(\tilde{\mathcal{N}}^{n,i}_{Y_t})) = 0$. In the contemporaneous graph the condition on neighbors blocks paths only if additionally paths through the parents of Y *and* all its neighbors are blocked. To this end, these parents need to be included in the conditioning set because (see Fig. 2.5) the motif $\rightarrow v -$ is open if node v is conditioned on. In Section 4.4 we further discuss estimation details and give an example.

In the physics literature also some attempts to reconstruct parents are discussed. For example, in Verdes (2005) a statistical procedure is described to infer the relevant most predictive sources. In Chapter 7 we will use the PC algorithm to infer optimal predictors for time series forecasting.

2.4.7. Underlying assumptions

Besides the basic assumption that we demand the multivariate process to possess an absolutely continuous joint probability density with respect to the product measure, the concept of conditional independence can only be interpreted causally with four main conditions: The Causal Markov Condition and Minimality were discussed in Sect. 2.4.3. The third one is *faithfulness* which guarantees that the graph entails all conditional independence relations true for the underlying process. Additionally, to call the links in the graph "causal" one assumes *causal sufficiency*, implying that no hidden common drivers are present (Spirtes et al., 2000). This assumption is

obviously violated if a finite set of, e.g., climate variables is analyzed (given the continuous nature of physical processes) and, as mentioned earlier, we call these links only "(Granger-) causal with respect to the variables taken into account".

As a counterexample that violates faithfulness consider the causal relation that the outcome of two independent fair coins X_1, X_2 influences a variable Y at a future time in the following way: If *both* coins simultaneously show heads or tails, $Y = 1$, and $Y = 0$ otherwise. Then $X_1 \perp\!\!\!\perp Y$ and $X_2 \perp\!\!\!\perp Y$ and the PC algorithm would converge assessing that both X_1 and X_2 are independent of Y. But actually the joint variable is not independent: $(X_1, X_2) \not\perp\!\!\!\perp Y$. Such a case could be covered by defining links not from a single process to Y, but from sets of processes. Another pathological example of unfaithfulness is a true graph with links $X \to Y \to Z$ and $X \to Z$ where the direct effect of X on Z is "counteracted" and fully balanced out through the mechanism via Y. Then $X \perp\!\!\!\perp Z$, but $X \not\perp\!\!\!\perp Z|Y$, which also would not be detected since the link between X and Z is already removed in the first step of the algorithm. But a counteracting mechanisms need not always fully erase another mechanism. Some examples of counteracting mechanisms are studied further in analytical examples in Chapter 5 and we also found this type of mechanism in real climate data as shown in Chapter 6.

The PC algorithm has the advantage that it is universally consistent (Spirtes et al., 2000). This is an important feature implying that the algorithm will actually converge to the true graph with probability 1 for $T \to \infty$ where T is the sample size. Unfortunately, no results about the rate of convergence exist, i.e., knowing at least how much samples we need for a given confidence. Uniform consistency would imply that the rate of convergence is independent of the true underlying graph, but there even is a proof that there *cannot be any* uniformly consistent causal discovery algorithm (Robins et al., 2003). Under certain assumptions on the distributions and the number of edges in the true graph, one can prove uniform convergence loosing the property of universality (Kalisch, 2007). Unfortunately, no such results exist for a more general class of processes.

To interpret a conditional independence relation between lagged processes causally, we need to be sure that the lag corresponds to an exact time lag, that is we assume the time points t to be without error. In particular in climate time series, while this holds for most climatological measurements in modern times, for time series of temperature and other variables from past climate this is not necessarily the case. Such variables are typically reconstructed from *proxy records* such as tree rings, speleothems growing in caves or marine sediments and a certain value has an often considerable uncertainty attached to it (Breitenbach et al., 2012). To be able to still distinguish cause and effect, the error must be smaller than the time sampling of the time series, which might additionally by irregular (Rehfeld et al., 2011; Rehfeld and Kurths, 2014).

2.5. Summary and epistemological aspects

In this chapter, we have briefly reviewed common approaches of measuring interactions in the physics and climate literature and discussed their weaknesses towards a causal interpretation. Our definition of causal interactions follows the idea of generalized Granger causality using the concept of conditional independence which encompasses general statistical associations. The central definition of time series graphs captures also the causal time lags and contemporaneous links which do not allow for a causal interpretation, but can nevertheless be important for a physical understanding of interactions. Finally, we presented an algorithm which allows to infer time series graphs efficiently avoiding the curse of dimensionality. The practical estimation of time series graphs will be studied in Chapter 4.

While Granger causality is well suited for stochastic processes, it cannot be well defined for deterministic dynamical systems which are better addressed with other approaches such as Sugihara et al. (2012) discussed in Sect. 2.3.4, or Daniusis et al. (2012); Janzing et al. (2012). As mentioned earlier, causality has also been phrased in a stricter sense than Granger causality to overcome the common critique "Correlation does not imply causation". With the idea of *interventions* and *causal calculus* (Pearl, 2000; Pearl, 2009), causation can be established for systems that can be experimentally manipulated. For example, if the barometer falls, the probability of rain is higher: The barometer Granger-causes rain. In the interventionalist's approach the barometer would be forcefully manipulated and it would be found that setting the barometer to low values does not cause rain. Conditional independence is then phrased using the concept of *do-calculus* (Pearl, 2000; Pearl, 2009) as

$$p(x, y|do(z)) = p(x|do(z)) \cdot p(y|do(z)) \quad \forall x, y, z \in \mathbb{R}^3, \tag{2.19}$$

where "do" actually implies an action that keeps the variable Z fixed at some value. In this way difficulties such as the one described above can be resolved. This approach can also be called an *active experiment*, while here we study the weaker 'non-manipulative', *passive* notion of Granger causality. Also, here we do not discuss approaches to infer from observations alone *whether* an unobserved latent process caused an interaction (Eichler, 2005). The framework of structural equation modeling here has the advantage that latent variables can be explicitly constructed and integrated in a model.

According to Bertrand Russell (1912): "The law of causation,[...] is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm". He claimed that the formulas of physics do not necessitate the notion of causality. Indeed, also the probabilistic notion of causality used in this thesis could be applied without this claim. But the notion of causality is a crucial explanatory concept in scientific research and it is important to see that Russel's critique does not address this explanatory function but rather a too demanding and hence implausible notion of causality. Thus, since we made clear that we use the term in this weaker probabilistic form of Granger causality and since it is an established term in the scientific literature, we stick to it.

Chapter 3.

Quantifying the strength of causal interactions

3.1. Introduction

3.1.1. Why quantifying causal interactions?

The previous chapter has established the statistical definition of causal directed and contemporaneous links in a multivariate process and in Chapter 4 it will be more practically shown how these can be estimated from time series. Now one could ask why the further step to also quantify the causal associations is necessary. After all, knowing the causal links is enough to construct a model and move on to simulations. Here we give some (partially overlapping) arguments why a quantification constitutes an important step in understanding a process:

1. **Ranking for model building:** Quantifying causal strength provides the modeler with a ranking of how important each link is and allows her to choose up to what *precision* she wants to model the process.

2. **Comparison:** From a descriptive perspective, a well-defined quantification allows to compare the strength of different coupling mechanisms in various aspects.

3. **Information efficiency:** To address the important question raised already in the introduction regarding global properties of information transfer and efficiency in a complex system (Latora and Marchiori, 2001), not only causal links, but also in a more general way, causal paths need to be quantified.

More from a machine learning perspective one might just be interested in

4. **Predictors:** From which nodes can a process best be predicted, which nodes carry the most information? In machine learning this relates to feature selection (Póczos et al., 2012).

5. **Exploratory data analysis:** From a data mining perspective, one might want to find important mechanisms in a huge data base where not every single mechanism can be analyzed by an expert (Reshef et al., 2011).

Each of these questions will demand different quantification measures. In model-based approaches as discussed in Section 2.3.2 these questions can be addressed by some cost function. But since the subject of interest here are complex systems, we will pursue this question from a more general perspective.

3.1.2. Properties for measures of multivariate dependence

Especially in the statistical literature many different measures of dependence exist which led already Rényi to formulate a set of axioms such a measure should fulfill (Rényi, 1959). Later Schweizer and Wolff (1981) updated this set of axioms with the most important ones being that it should be larger than zero and zero if and only if the variables are independent. Further, the measure should be invariant to strictly monotonous transformations such as arise from rescaling the units of measurement of a variable (Rényi even demanded invariance to one-to-one transformations). Recently, Reshef et al. (2011) put forward two more heuristic demands such a measure should fulfill:

1. **Generality:** The measure should not be restricted to certain types of associations like linear measures.

2. **Equitability:** The measure should reflect a certain heuristic notion of coupling strength, i.e., it should give similar scores to equally noisy dependencies. The latter is especially important for comparisons and ranking of the strength of dependencies.

However, Reshef et al. (2011) and Schweizer and Wolff (1981) refer only to pairs of random variables. For the multivariate case we propose to add four properties:

3. **Lag-specific:** The measure should quantify dependence between *lagged* time series. This is important for physical interpretations.

4. **Causality:** As defined in the previous chapter, the measure should give a non-zero value only to the dependency between lagged components of a multivariate process that are not independent conditional on the remaining process.

5. **Coupling strength autonomy:** Also for dependent components we seek for a causal notion of coupling strength that is well interpretable, in that it is uniquely determined by the interaction of the two or more components alone and in a way autonomous of their interaction with the remaining processes.

6. **Practical computability:** The measure should not be defined using infinite dimensionality that would require arbitrary truncations.

The property of causality will also be used in a wider sense to incorporate not only causal links, but also causal paths. To understand the property of coupling strength autonomy, consider a simple example: Suppose we have two interacting processes X and Y and a third process Z, that drives both of them. Then a bivariate measure of

coupling strength between X and Y will be influenced by the common input of Z, while our demand is, that the measure should be autonomous of the interactions of X and Y with Z. Using do-calculus, in the experimental setting this corresponds to keeping Z fixed and solely measuring the impact of a change in X on Y averaged over all realizations of Z. In a non-interventionalist setting, it corresponds to observing the effect of changes in X on Y for different Z. This property can be regarded as one ingredient of a multivariate extension of the equitability property and can also be generalized to more than two processes as will be shown in this chapter and Chapter 5.

While Rényi's strong axioms are only fulfilled by the *maximum correlation coefficient* (Gebelein, 1941) and the *Randomized Dependence Coefficient* (Lopez-Paz et al., 2013) (which is actually practically computable), the axioms by Schweizer and Wolff (1981) are fulfilled by information theoretic measures like Rényi and Tsallis divergences (Póczos et al., 2012) and, therefore, also by the special case of Shannon type mutual information (MI) and conditional mutual information (CMI) defined in Sect. 3.2.2. The axioms in Schweizer and Wolff (1981) also express the idea of generality. Additionally to generality, the authors in Reshef et al. (2011) demonstrate that their algorithmically motivated *maximal information coefficient* fulfills the property of equitability. We will discuss this property in more detail in Section 4.2.5. However, apart from issues with statistical power, a crucial drawback of their measure as well as the one by Lopez-Paz et al. (2013) is, that it is not clear how to extend it to the conditional case that a multivariate causal analysis demands. Another very recent measure is the conditional distance correlation (Póczos and Schneider, 2012) which might be promising, but demands the prior estimation of densities and is not further discussed here.

There are few works considering a concept of coupling strength in the multivariate context of causality. In Jachan et al. (2009); Schelter et al. (2009) this problem is approached in the linear framework of partial directed coherence and in Chen et al. (2004); Marinazzo et al. (2008) using the less restricted, yet still model-based, concept of Granger causality, all sharing the problem that the model might be misspecified.

In Ay and Polani (2008) and Janzing et al. (2013) an idea is described that is most similar to our approach in that there the question of quantifying links is seen as a second step based on the known directed acyclic graph. Ay and Polani (2008) address the problem from an interventionalist perspective using Pearl's do-calculus (Pearl, 2000) which we do not further discuss here since we assume the process to be not manipulable. Janzing et al. (2013) also use an information-theoretic approach, but based on a different set of postulates. We discuss their idea in the context of communication theory in Sect. 5.5.1.

3.1.3. The idea of momentary information

Our approach to a measure of a causal coupling strength formally introduced in Sect. 3.4.5 is based on the fundamental concept of *source entropy*, also termed the *entropy rate* (Shannon, 1948; Shannon and Weaver, 1963), and for the special case

of bivariate ordinal pattern time series the *momentary information transfer* (MIT) has been introduced in Pompe and Runge (2011). Consider a symbol-generating process X. At each time t a realization x_t is generated. Now the source entropy of X_t measures the uncertainty about x_t before its observation if all former symbols $(x_{t-1}, x_{t-2}, \ldots)$ are known (entopies will be formally introduced in Sect. 3.2). For a completely deterministic non-chaotic system the source entropy will always be zero, but for a real world process there will always be some uncertainty stemming from *dynamical noise*. This type of noise is to be distinguished from *observational noise* which usually contaminates each measured time series (Schreiber and Kantz, 1995), but has no effect on the dynamics of the process. Dynamical noise might occur due to unresolved smaller-scale processes and can be modeled by adding a random variable to the system. More formally, consider a subprocess X of a multivariate process $\mathbf{X}$, that is described by the discrete-time equation

$$X_t = f\left(Z^1_{t-\tau_1}, Z^2_{t-\tau_2}, \ldots\right) + \eta^X_t, \tag{3.1}$$

with some arbitrary function f of the other subprocesses at past times $Z^1_{t-\tau_1}$, $Z^2_{t-\tau_2}$, $\ldots \in \mathbf{X}^-_t$ and the random part subsumed under η^X_t. Note that the noise could also occur in a multiplicative part. Here the uncertainty of an outcome x_t will *on average* be reduced if a realization of the past $Z^1_{t-\tau_1}$, $Z^2_{t-\tau_2}$, $\ldots$ is known. But for non-zero η^X_t there will always be some "surprise" left when observing x_t. This surprise gives us information and the expected information here is the source entropy $H(X_t|\mathbf{X}^-_t) = H(\eta^X_t)$ of X. Due to measurement errors ϵ, we will in general not be able to estimate the source entropy alone, but only $H(X_t + \epsilon^X_t|\mathbf{X}^-_t + \epsilon^{\mathbf{X}^-}_t)$. Even assuming a perfect measurement apparatus for a deterministic dynamical system without dynamical noise, the entropy rate h^{symb} – since it is computed by creating a symbol sequence from a coarse graining in phase-space – depends on some resolution parameter r. Then the limit $\lim_{r\to 0} h^{\text{symb}}$ might exist and is then called the *Kolmogorov-Sinai entropy*. If this limit is finite and larger than zero, the system is called chaotic. But here we study stochastic and also discrete time processes because the finite set of measured variables of a complex system like the Earth will never perfectly describe the full system's state and all remaining processes contribute to dynamical noise (implying that the Kolmogorov-Sinai entropy diverges).

The momentary information entering at each time t in a subprocess X can also be understood as continuous (statistically phrased: stationary) perturbations. The measures defined in this chapter are intended to quantify how these perturbations propagate in the complex system. With this central idea we define several measures that allow to quantify the interaction between two causally linked processes (*momentary information transfer* (MIT)), but also along causal paths and between multiple processes. In Chapter 5, we mathematically prove for a very general class of systems that this class of measures is practically computable and fulfills the properties of generality, causality and coupling strength autonomy, while the more complex property of equitability will be addressed in Section 4.2.5. In Chapter 5, we also discuss how these measures provide a way to quantify more global properties of

complex systems such as the efficiency of information transfer and study how causal paths can be analyzed. Because the linear framework has clear advantages over non-parametric methods (as will be demonstrated in Chapter 4) given limited data, in this chapter we explore the information-theoretic approach alongside with the simplest model-based linear partial correlation – dropping the property of generality – for which the same idea of momentary information transfer can be implemented, then better described as *momentary variance transfer*.

This chapter serves to introduce the measures to quantify causal strength which are in detail studied in analytical and numerical examples in Chapter 5. To this end, we first review basic concepts of information theory (Sect. 3.2) and the linear theory of partial correlation and regression (Sect. 3.3). Then, in Sect. 3.4 we define time series based interaction measures from the well-known lagged mutual information and transfer entropy to the novel measures based on momentary information as published in Pompe and Runge (2011); Runge et al. (2012a); Runge et al. (2012b). Extending this approach, in Sect. 3.5 we define measures to quantify interactions between multiple processes.

3.2. Information theory

3.2.1. Entropy and conditional entropy

In his theory of communication, Shannon investigated how efficient information can be transferred over a channel in the presence of noise. To this end he introduced a statistical notion of *entropy* as a central concept of his information theory. Inspired by the thermodynamic entropy introduced by Clausius (1862); Boltzmann (1872); Gibbs (1902), $S = k_b \sum p_i \log p_i$ (where k_b is the Boltzmann constant and p_i the probability of a system's state), Shannon's entropy describes the uncertainty about the transmission of a message over a noisy channel. Lindley (1956) introduced the view of a statistical sample as an example of Shannon's noisy channels that contains a message about parameters (Soofi, 1994; Golan, 2002). Then the Shannon entropy quantifies the uncertainty about the outcome of the "experiment" of measuring a random variable, i.e., its "randomness". The probability of an outcome p_i carries the information $-\log p_i$. The more rare an outcome is, the higher its information value. The entropy is the expectation value over all outcomes. While the long standing discussion between information theory and the foundations of statistical mechanics (Jaynes, 1957; Crutchfield and Shalizi, 1999; Allahverdyan et al., 2009) is beyond the scope of this work, we will give a modest approach to interpret the information-theoretic measures introduced in this thesis thermodynamically in Sect. 5.5.2.

For a continuous random variable X with density function $p_X(x)$, which we will abbreviate to $p(x)$, Shannon defined[4] the *continuous or differential entropy* as

$$H(X) \equiv -\int p(x) \ln p(x)\, dx, \tag{3.2}$$

where the integral runs over the support of the density, i.e., the values for which the density is larger than zero. The convention is that $0 \ln(0) = 0$. Here we use the natural logarithm as the basis and all information theoretic quantities will, therefore, be measured in *nats*.

The *multivariate entropy* (or *joint entropy*) is analogously defined as

$$H(X_1, \ldots, X_N) \equiv -\int \cdots \int p(x_1, \ldots, x_N) \ln p(x_1, \ldots, x_N)\, dx_1 \cdots dx_N. \tag{3.3}$$

An important property of the Shannon entropy[5] is the chain rule

$$H(X, Y) = H(X) + H(Y|X), \tag{3.4}$$

from which the *conditional entropy*

$$H(Y|X) \equiv H(X, Y) - H(X) \tag{3.5}$$

$$= \int p(x) H(Y|X = x)\, dx \tag{3.6}$$

can be defined. Figure 3.1 illustrates the conditional entropy as represented in Eq. (3.6) as a scalar functional of the conditional probability density.

Discrete and continuous entropies The Shannon entropy for discrete random variables can analogously be defined where the summation over the possible alphabet replaces the integral. Here, we stick to the continuous case because the geophysical variables analyzed in the applications are typically continuous and the estimators introduced in Sect. 4.2 are based on the continuous case. Nevertheless, the theory developed in this thesis holds equivalently for the discrete case. One important difference is that the discrete entropy measures randomness in an absolute way, while the continuous entropy measures randomness relative to the coordinate system and can become negative (Shannon, 1948). In fact the entropy in new coordinates x' is

[4] Equation (3.2) is not the only way to define an entropy. Alfréd Rényi (1961) generalized Shannon's definition to the class of Rényi entropies of order α:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\int p(x)^\alpha \right).$$

For the limiting case $\alpha \to 1$ the Shannon type entropy is recovered.

[5] Contrary to the Rényi entropies for $\alpha \neq 1$.

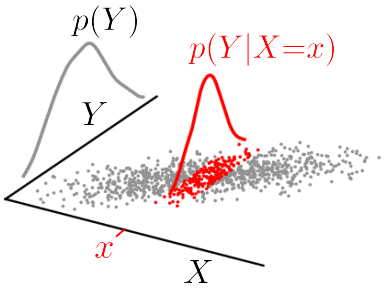


Figure 3.1.: Scatter plot of conditional probability. The conditional entropy $H(Y|X)$ given by Eq. (3.6) is a weighted integral (conditional expectation) over all entropies $H(Y|X=x)$ at a particular value x.

given by

$$H(X') = H(X) - \int p(x) \ln J(x, x')\, dx, \tag{3.7}$$

where $J(x, x')$ is the Jacobian of the coordinate transformation. That is, the continuous entropy measures randomness relative to the coordinate system with each volume element given equal weight where $\ln J(x, x') = 0$. For $x' = x + c$ with some constant c, the Jacobian is one and it follows that entropy is *translationally invariant*. Note that differential entropy can also be negative, however the concept of mutual information introduced in the next section depends on a difference of entropies which makes it invariant under coordinate transformations and non-negative.

3.2.2. Mutual information and conditional mutual information

While entropy is a measure of the uncertainty about outcomes of one process, *mutual information* (MI) is a measure of the reduction thereof if another process is known. The Shannon type[6] MI is defined as

$$I(X;Y) \equiv \int\int p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}\, dxdy \tag{3.8}$$

$$= H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{3.9}$$

$$= H(X) + H(Y) - H(X,Y), \tag{3.10}$$

[6] One cannot in general well define a Rényi mutual information. However, in Pompe (1993) this has been done for the case of $\alpha = 2$. A practical computational advantage of Rényi entropies and mutual information is that the logarithm is taken *after* the summation which allows for faster algorithms.

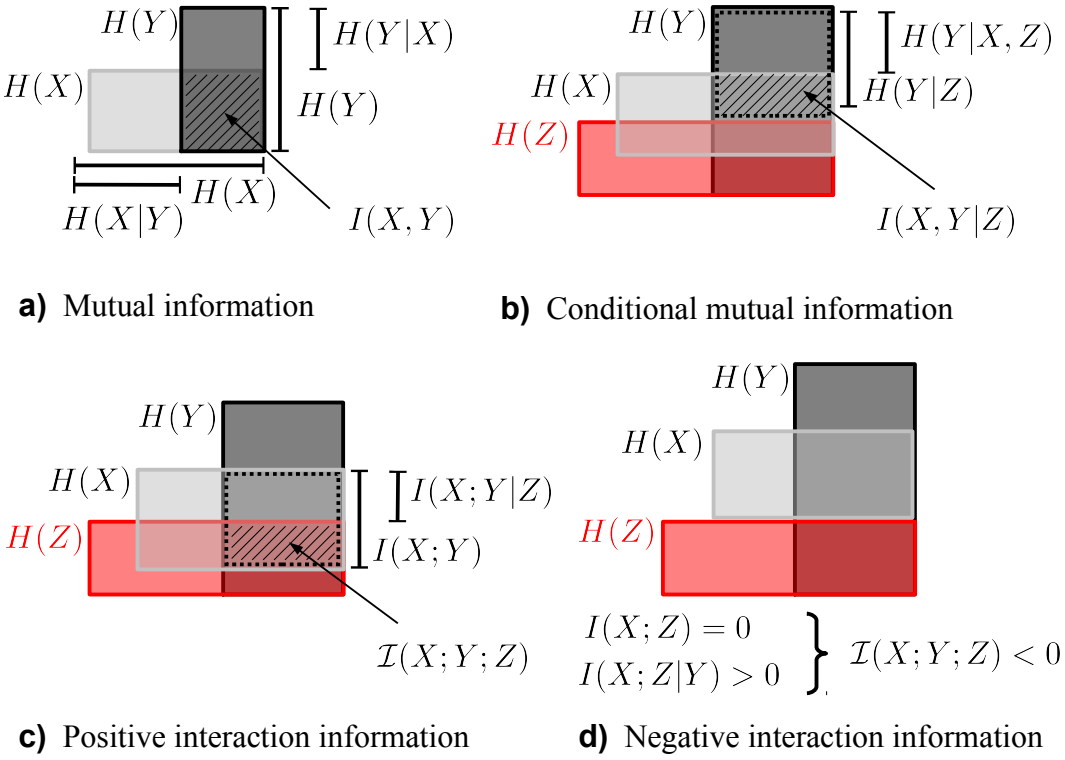


Figure 3.2.: Venn diagrams of (conditional) mutual information and interaction information. Note that the analogy of entropies to sets as suggested by Venn diagrams should not be exaggerated since the interaction information can also be negative (for example, if the entropies of X and Z do not 'overlap' anymore as shown in (d)).

i.e., in the form of Eq. (3.9) as the difference between the uncertainty in Y and the remaining uncertainty if X is already known (and vice versa). Or in Eq. (3.8) as a certain Kullback-Leibler distance (Kullback and Leibler, 1951; Cover and Thomas, 2006) between the distributions $p(x, y)$ and the distribution for the independent case $p(x)p(y)$. From the definition it immediately follows that MI is symmetric in its arguments

$$I(X;Y) = I(Y;X) \tag{3.11}$$

and zero if and only if X and Y are independent

$$I(X;Y) = 0 \quad \Leftrightarrow \quad p(x,y) = p(x)p(y) \quad \forall x, y \in \mathbb{R}^2. \tag{3.12}$$

Further, using Jensen's inequality (Cover and Thomas, 2006) one can show that MI is always non-negative (which holds for the discrete as well as the continuous case),

$$I(X;Y) \geq 0, \tag{3.13}$$

from which it also follows for the joint and conditional entropies that

$$H(X,Y) \leq H(X) + H(Y) \quad \Leftrightarrow \quad H(Y|X) \leq H(Y), \tag{3.14}$$

i.e., conditioning can only reduce the uncertainty and $H(X,Y) = H(X) + H(Y)$ only if X and Y are independent. In this case the uncertainty about the pair (X,Y) is, therefore, the sum of the single uncertainties.

The most important measure used throughout this thesis is the *conditional mutual information* (CMI) given by

$$\begin{aligned} I(X;Y|Z) &\equiv H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) && (3.15)\\ &= \int p(z) \iint p(x,y|z) \log \frac{p(x,y|z)}{p(x|z) \cdot p(y|z)} \, dxdydz && (3.16)\\ &= \int p(z) I(X;Y|Z=z) \, dz && (3.17)\\ &= H(Y|Z) - H(Y|X,Z) = H(X|Z) - H(X|Y,Z) && (3.18)\\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z). && (3.19) \end{aligned}$$

It can be phrased as the mutual information between X and Y that is not contained in a third variable Z. Just like MI, CMI is non-negative and symmetric $I(X;Y|Z) = I(Y;X|Z)$. CMI (and MI) is bounded by the marginal (conditional) entropies

$$0 \leq I(X;Y|Z) \leq \min(H(X|Z),\, H(Y|Z)). \tag{3.20}$$

Further, CMI is zero if and only if X and Y are independent *conditionally on* Z,

$$X \perp\!\!\!\perp Y|Z \quad \Longleftrightarrow \quad I(X;Y|Z) = 0. \tag{3.21}$$

This property makes CMI especially useful to measure conditional independence as needed to estimate time series graphs. Figures 3.2(a) and (b) visualize MI and CMI in Venn diagrams as a difference of conditional entropies. In this representation also the symmetry in the arguments is obvious.

Some further properties are important later on. The random variables X, Y, Z can also be multivariate and we will sometimes use multivariate CMIs like $I((X,W);Y|Z)$ where the colon always separates the arguments. For multivariate CMIs also a *chain rule* holds which we frequently utilize in this thesis:

$$\begin{aligned} I(X_1,\ldots,X_N;Y|Z) &= I(X_1;Y|Z) + I(X_2,\ldots,X_N;Y|X_1,Z)\\ &\vdots \end{aligned}$$

$$= \sum_{i=1}^{N} I(X_i; Y| \cup_{j=1}^{i-1} X_j, Z). \tag{3.22}$$

The data processing inequality (Cover and Thomas, 2006) states that

$$I(X; f(Y)|Z) \leq I(X; Y|Z), \tag{3.23}$$

i.e., manipulating Y by some function f can only reduce the shared information. Note, however, that CMI is invariant under smooth uniquely invertible transformations such as linear rescalings of X, Y or Z. This can easily be seen from Eq. (3.7), because such transformations alter the joint density to

$$p_{X'Y'Z'}(x', y', z') = J(x, x')J(y, y')J(z, z')p_{XYZ}(x, y, z)\,, \tag{3.24}$$

where $J(\cdot, \cdot)$ denotes the Jacobians. Correspondingly, the marginal entropies are altered and in the formula for CMI (Eq. (3.15)) the second terms in Eq. (3.7) then cancel out. For random variables Y and W and an arbitrary function f we have that

$$\begin{aligned} H(Y + f(W)|W) &= \int p(w)H(Y + f(W)|W = w)dw \\ &= \int p(w)H(Y|W = w)dw \\ &= H(Y|W), \end{aligned} \tag{3.25}$$

because $f(W)$ for $W = w$ is a fixed constant and entropies are translationally invariant. In particular, $H(f(W)|W) = 0$. This property also holds for the joint entropy and it follows for CMI that

$$I(X + g(Z); Y + f(W)|Z, W) = I(X; Y|Z, W). \tag{3.26}$$

Also here, $I(X; f(W)|W) = 0$. Last, conditions that are conditionally independent of X and Y given Z can be dropped:

$$X \perp\!\!\!\perp W|Z \text{ and } Y \perp\!\!\!\perp W|Z \implies I(X; Y|Z, W) = I(X; Y|Z), \tag{3.27}$$

which can be easily derived from $I((X, Y); W|Z) = 0 \Rightarrow H(X, Y|Z, W) = H(X, Y|Z)$ and correspondingly for the marginals. With these properties we will be able to analytically derive the measures to be introduced now for nonlinear stochastic processes in Chapter 5.

3.2.3. Interaction information

Just as MI and CMI are differences of conditional entropies, also the difference of CMIs has an interesting interpretation that we will utilize to measure the effect of one random variable on the interaction between two others. Such a measure has been studied by Abramson (1963); Tsujishita (1995); Leydesdorff and Sun (2009) under

the name *multiple information*. We use the term *interaction information* with the symbol $\mathcal{I}$ defined as

$$\begin{aligned}\mathcal{I}(X;Y;Z) &\equiv I(X;Y) - I(X;Y|Z) \\ &\equiv I(Y;Z) - I(Y;Z|X) \\ &\equiv I(Z;X) - I(Z;X|Y).\end{aligned} \tag{3.28}$$

In McGill (1954); Jakulin and Bratko (2003) this quantity, denoted by $I(X;Y;Z)$, is defined with the signs reversed, but our definition is more consistent with the definition of mutual information [Eq. (3.9)], i.e., the conditional quantity is subtracted from the unconditional one. It is also straightforward to define the *conditional interaction information*

$$\mathcal{I}(X;Y;Z|W) \equiv I(X;Y|W) - I(X;Y|Z,W). \tag{3.29}$$

Contrary to CMI, the (conditional) interaction information can also be negative and is bounded by

$$\begin{aligned}&-\min(I(X;Y|Z,W), I(Y;Z|X,W), I(Z;X|Y,W)) \\ &\leq \mathcal{I}(X;Y;Z|W) \\ &\leq \min(I(X;Y|W), I(Y;Z|W), I(Z;X|W)).\end{aligned} \tag{3.30}$$

The possible negativity also shows that the visualization in Fig. 3.2(c) as sets in Venn diagrams should not be overinterpreted. In Fig. 3.2(d) a case is shown where X and Z are *unconditionally* independent, but conditionally dependent leading to $I(X;Z|Y) \geq I(X;Z)$ and, therefore, a negative interaction information. That this property can actually by intuitively understood will be studied in examples in Sect. 5.2.4.

3.3. Linear theory

As mentioned before, we develop the formalism of quantifying causal strength in parallel with linear measures. Here, we briefly review concepts from the linear theory of regressions and partial correlation.

3.3.1. Regression

In regression analysis, the influence of possibly multiple variables, the *regressors* $\mathbf{U}$, on (the mean of) a *dependent* variable Y is estimated. In the most common multiple *linear* regression using the model

$$Y = \mathbf{U}\mathbf{B} + \varepsilon_Y, \tag{3.31}$$

with residual error term ε_Y and where the regression coefficient vector $\mathbf{B}$ is – assuming ε_Y and $\mathbf{U}$ are independent – given by

$$\mathbf{B} = \Gamma_{\mathbf{U}}^{-1}\Gamma_{\mathbf{U};Y}, \tag{3.32}$$

where $\Gamma_{\mathbf{U}} \equiv E[\mathbf{U}^\top\mathbf{U}]$ is the covariance of the regressors and $\Gamma_{\mathbf{U};Y} \equiv E[\mathbf{U}^\top Y]$ the vector of covariances of each regressor with the dependent variable Y ($E[\cdots]$ denotes the expectation). The assumptions entering such an approach are that the regressors $\mathbf{U}$ are without measurement error, heteroscedasticity (constant variance of ε_Y), independence of errors, lack of collinearity of the regressors and, of course, that Y is a linear combination of the regressors. Some assumptions can be relaxed using more sophisticated estimation techniques (weighted regression, errors-in-variable models, etc.). Most important, though, is the assumption of linearity, which means that if the model correctly describes the observed process, the coefficients $\mathbf{B}$ are the physically interesting parameters.

3.3.2. Partial correlation

Cross correlation and partial correlation are the linear counterparts to mutual information and conditional mutual information. The cross correlation of zero-mean random variables X, Y is given by

$$\rho(X;Y) \equiv \frac{E[X^\top Y]}{\sqrt{E[Y^\top Y]}\sqrt{E[X^\top X]}}, \tag{3.33}$$

which depends on the covariances and variances. If one regresses two variables X, Y on the same regressors $\mathbf{U}$, then the cross correlation between the residuals

$$\begin{aligned} X_{\mathbf{U}} &= X - \mathbf{U}\Gamma_{\mathbf{U}}^{-1}\Gamma_{\mathbf{U};X} \\ Y_{\mathbf{U}} &= Y - \mathbf{U}\underbrace{\Gamma_{\mathbf{U}}^{-1}\Gamma_{\mathbf{U};Y}}_{\substack{\text{regression}\\ \text{coefficient}\\ \text{(vector)}}} \end{aligned} \tag{3.34}$$

is the *partial correlation*

$$\rho(X;Y|\mathbf{U}) = \rho(X_{\mathbf{U}};\, Y_{\mathbf{U}}). \tag{3.35}$$

This is also the way in which we estimate partial correlation as numerically studied in Chapter 4.

Difference between partial correlation and CMI CMI is always non-negative, the partial correlation, on the other hand, can be negative, which can be interpreted as the variables X and Y being anticorrelated (i.e., an increase in X is related to a decrease in Y). The notion of an anticorrelation already implies that a certain model

is imposed by which we measure the relationship between X and Y and can serve as an important information to model an interaction. Partial correlation captures only the linear part, i.e., the first two moments, of an association. In fact, if the process is a multivariate Gaussian and is therefore perfectly described by the first and second moments, the conditional mutual information is a function of the partial correlation:

$$I(X;Y|\mathbf{U}) = -\frac{1}{2}\ln\left(1-\rho(X;Y|\mathbf{U})^2\right). \tag{3.36}$$

In Chapter 4, we will see that partial correlation has immense practical advantages over information theory such as no bias and low variance even for very small sample sizes. The novel measures introduced in the next section can equally be based on (conditional) mutual information or partial correlation, for which the possible negativity needs to be taken into account.

3.4. Time series (graph)-based measures of dependence between two processes

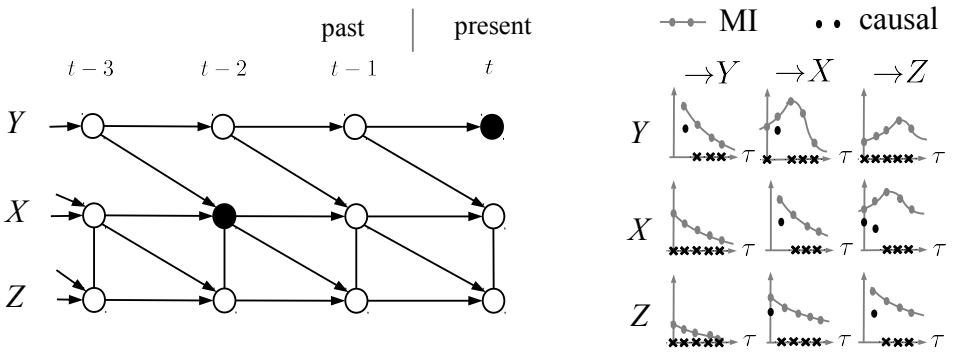
After the preceding sections, we are now equipped with measures to quantify interactions between components in a multivariate process. We will see how the combination of information-theoretic measures with the concept of time series graphs as introduced in Chapter 2 allows to quantify causal interactions. That is, the determination of the strength and delay of a mechanism now is a two-step procedure. In the first step, the time series graph is estimated which determines the existence or absence of a link and thus of a (Granger) causality between lagged components the multivariate process. For the second step discussed now, we will introduce several measures based on (conditional) mutual information that quantify the interaction between two components. But first, we review common measures such as mutual information and transfer entropy (Schreiber, 2000). For all measures, the linear counterpart is obtained by replacing the (conditional) mutual information with the (partial) correlation, i.e., the "I" with the "ρ".

3.4.1. Lagged mutual information

The first and simplest association measure combining information theory with time series is the lagged (cross-)mutual information given by

$$I^{\mathrm{MI}}_{XY}(\tau) \equiv I(X_{t-\tau};Y_t). \tag{3.37}$$

For $\tau > 0$, MI measures the information in the past of X that is contained in Y. Contrary to the common plot of lag functions against positive and negative lags τ shown in Fig. 2.1, the presentation as a matrix of lag functions in Fig. 3.3(b) with only non-negative lags underlines the interpretation of the lag functions (of mutual information or cross correlation) as directional influences. In analogy, the auto-MI is defined as $I(Y_{t-\tau};Y_t)$ for $\tau > 0$ (for $\tau = 0$ this is the entropy $H(Y)$).



a) Time series graph **b)** Lag functions

Figure 3.3.: Visualization of causal interactions in a multivariate process $\mathbf{X}$ via (a) a time series graph which depicts the causal links and (b) lag functions which quantify lagged associations. The lagged mutual information is shown in grey and the causal associations are marked by black dots (crosses mark non-causal associations). For example, in the center left panel the lag function shows the value of the MI between $X_{t-\tau}$ and Y_t for $\tau \geq 0$. The case $\tau = 2$ is marked by the black dots in (a). Because these two processes share common information from (unblocked paths coming from) Y_{t-3} and further in the past, the MI is non-zero, even though there is no causal link between the two.

Cross-MI is not intended to exclude entropy common to both $X_{t-\tau}$ and Y_t, yet it is frequently used to determine the time delay of an interaction mechanism, admitting that there does not necessarily exist a causal relation. In this way, also the linear version, cross correlation, is widely used in climate research. That this use is prone to pitfalls even for the modest goal of determining a coupling delay is extensively discussed in Sect. 5.2.1. The ambiguity in interpreting the value of MI is discussed in Sect. 5.2.2 and severe weaknesses regarding an assessment of significance for time series with autocorrelations typically occurring in climate are studied in Sect. 4.3.

3.4.2. (Decomposed) transfer entropy

As listed in Chapter 2, towards a causal interpretation, measures need to be able to exclude information from the common past. Implementing this idea, Schreiber introduced *transfer entropy* (TE) (Schreiber, 2000) which is the information-theoretic analogue of Granger causality and for multivariate Gaussian processes they can actually be shown to be equivalent (Barnett et al., 2009). In the original article TE was defined between two variables, but here we will focus on the more general multivariate case that admits to exclude also influences from other processes. This section is based on results published in Runge et al. (2012a)[7].

[7]Joint work with substantial contributions by co-author J. Heitzig.

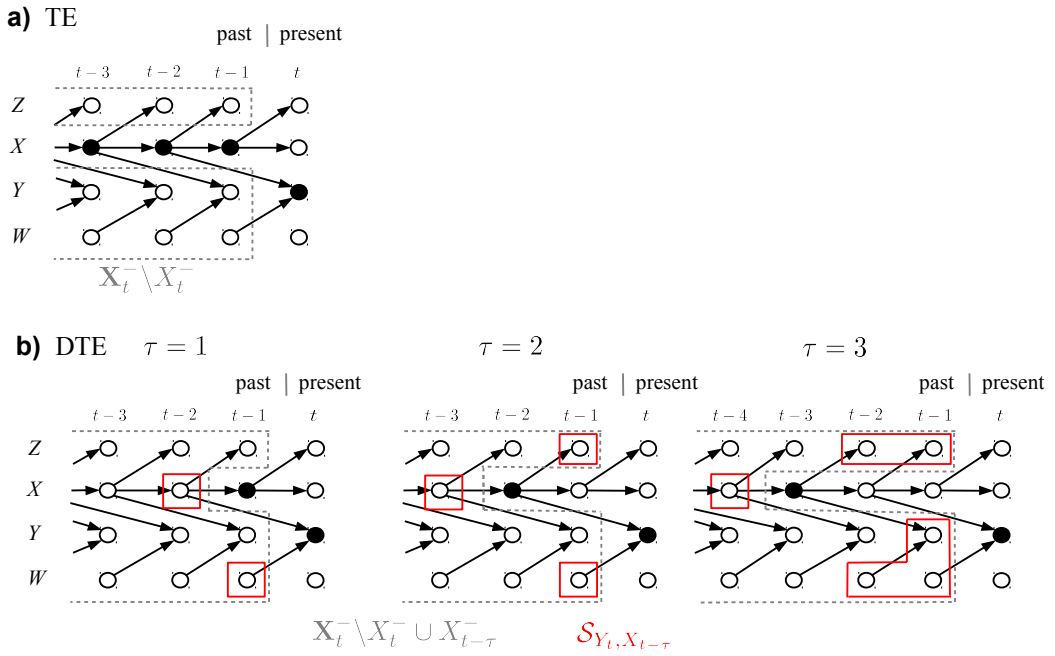


Figure 3.4.: TE and DTE for a multivariate example process as given by Eq. (5.3) that will be analytically analyzed in Sect. 5.2.2. (a) TE given by Eq. (3.38) between the infinite past vector X_t^- and Y_t (black dots) conditioned on the remaining infinite past $\mathbf{X}_t^- \backslash X_t^-$ (gray dashed open box). (b) First three summands of DTE given by Eq. (3.42). For the CMI between $X_{t-\tau}$ and Y_t (black dots) only the finite set $\mathcal{S}_{Y_t, X_{t-\tau}}$ (red solid boxes) is needed to satisfy the Markov property (Eq. (3.41)). $\mathcal{S}_{Y_t, X_{t-\tau}} \subset \mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-$ (gray dashed open box) must be chosen so that it *separates* the remaining infinite conditions $(\mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-) \backslash \mathcal{S}_{Y_t, X_{t-\tau}}$ from Y_t in the graph (for a definition of paths and separation see section 2.4.3).

Given a stationary multivariate discrete-time stochastic process $\mathbf{X}$, we denote its uni- or multivariate subprocesses $X, Y, Z, W, \ldots$ and the random variables at time t as $\mathbf{X}_t, X_t, \ldots$. Their *pasts* are defined as $\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots)$ and $X_t^- = (X_{t-1}, X_{t-2}, \ldots)$. Now TE [see Fig. 3.4(a)]

$$I^{\mathrm{TE}}_{X \to Y} \equiv I(X_t^-; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-) \tag{3.38}$$

is the reduction in uncertainty about Y_t when learning the past of X_t, if the rest of the past of $\mathbf{X}_t$, given by $\mathbf{X}_t^- \backslash X_t^-$, is already known (where "\" denotes the subtraction of a set). Note that, because of the assumed stationarity, $I^{\mathrm{TE}}_{X \to Y}$ is independent of t. TE measures the aggregated influence of X at all past lags and is *not* lag-specific. The definition of TE leads to the problem that infinite-dimensional densities have to be estimated, which is commonly called the "curse of dimensionality" mentioned in the introduction. In the usual naive estimation of TE the infinite vectors are simply

truncated at some $\tau_{\max}$ leading to

$$I_{X\to Y}^{\mathrm{TE},\tau_{\max}} \equiv I(X_t^{(t-1,\ldots,t-\tau_{\max})}; Y_t \,|\, \mathbf{X}_t^{(t-1,\ldots,t-\tau_{\max})} \backslash X_t^-). \tag{3.39}$$

where $X_t^{(t-1,\ldots,t-\tau_{\max})} = (X_{t-1}, \ldots, X_{t-\tau_{\max}})$ (correspondingly for $\mathbf{X}$) and $\tau_{\max}$ has to be chosen at least as large as the maximal coupling delay between X and Y, which can lead to very large dimensions. In our numerical experiments in Section 4.2 we will demonstrate that the choice of a truncation lag $\tau_{\max}$, which affects the estimation dimension via $D = N \cdot \tau_{\max} + 1$ (where N is the number of processes), has a strong influence on the value of TE and affects the reliability of causal inference. This is a huge disadvantage because the coupling delay should not have an influence on the measured coupling strength. This severe limitation can be overcome by embedding TE into the framework of time series graphs as follows. There are two infinite-dimensional parts in TE: X_t^- and $\mathbf{X}_t^- \backslash X_t^-$. We address the first by decomposing TE into contributions of individual lags of X via the chain rule (Eq. (3.22), for detailed derivations see Appendix A.1),

$$I(X_t^-; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-) = \sum_{\tau=1}^{\infty} I(X_{t-\tau}; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-, X_{t-\tau}^-)\,. \tag{3.40}$$

Now the decisive step to escape the still infinite dimension of the condition in each term is done by utilizing the knowledge of the Markov property Eq. (2.12) encoded in the time series graph. It implies that most conditions can be dropped utilizing Eq. (3.27),

$$I(X_{t-\tau}; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-, X_{t-\tau}^-) = I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}), \tag{3.41}$$

for a certain *finite* subset $\mathcal{S}_{Y_t,X_{t-\tau}} \subset \mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-$ of the conditions (see Fig. 3.4 (b)). These sets $\mathcal{S}_{Y_t,X_{t-\tau}}$ can be determined *after* the time series graph has been inferred and TE can be estimated using only low-dimensional densities. The remaining infinite sum can be truncated at some finite τ^* since the terms typically decay exponentially with τ:

$$I_{X\to Y}^{\mathrm{TE}} \approx I_{X\to Y}^{\mathrm{DTE}} = \sum_{\tau=1}^{\tau^\star} I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t,X_{t-\tau}}) \tag{3.42}$$

with $\tau^\star$ chosen as the smallest τ for which the estimated remainder is smaller than some given absolute tolerance (see Appendix A.1 for details). This can improve the estimation of TE considerably as compared to the direct estimation as shown in Sect. 5.4, although the sets $\mathcal{S}_{Y_t,X_{t-\tau}}$ and the resulting dimensions can still be large.

The summands in Eq. (3.42) can be seen as the contributions of different lags to TE, but should not be interpreted as lag-specific causal contributions because they can be non-zero also for lags τ for which there is no link in the graph. Finally, apart from the issue of high dimensionality and lag-specific causality, we will demonstrate

in Sect. 5.2.2 and Sect. 5.4 that TE or DTE can be rather counter-intuitive and misleading as measures of coupling strength.

3.4.3. Link-defining conditional mutual information

As mentioned earlier, conditional mutual information can be used to measure conditional independence via Eq. (3.21). In this way, the conditional independence used to define links in the time series graph (Eq. (2.8)) could be estimated by a certain CMI as introduced in Runge et al. (2012b). For $\tau > 0$ the CMI

$$I^{\text{LINK}}_{X\to Y}(\tau) \equiv I(X_{t-\tau}; Y_t|\mathbf{X}_t^- \setminus \{X_{t-\tau}\}) \tag{3.43}$$

defines a directed link "$X_{t-\tau} \to Y_t$" and for $\tau = 0$ the CMI

$$I^{\text{LINK}}_{X-Y} \equiv I(X_t; Y_t \,|\, \mathbf{X}^-_{t+1}\backslash\{X_t, Y_t\}) \tag{3.44}$$

defines an undirected contemporaneous link where also the contemporaneous present $\mathbf{X}_t\backslash\{X_t, Y_t\}$ is included in the condition. Like TE, these CMIs involve infinite-dimensional vectors and can thus not be directly computed, but only involving truncations. As shown in Sect. 5.4, this measure therefore suffers from the problem of high dimensionality and also theoretically does not fulfill the coupling strength autonomy property as analyzed in Sect. 5.2.2.

3.4.4. Information transfer

As the Markov property given by theorem Eq. (2.12) implies that links as defined through $I^{\text{LINK}}_{X\to Y} > 0$ can equivalently be defined by conditioning only on the parents of Y_t (see Eq. (3.27)), we introduce, following Runge et al. (2012b), the *information transfer to* Y (ITY)

$$\begin{aligned} I^{\text{ITY}}_{X\to Y}(\tau) &\equiv I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}) &(3.45)\\ &= H(Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}) - H(Y_t|\mathcal{P}_{Y_t}), &(3.46)\end{aligned}$$

where the latter term is the source entropy of Y. ITY is non-zero only for dependent nodes (and therefore fulfills the properties of generality and causality) and used in the algorithm to estimate the time series graph discussed in Chapter 2. A similar measure has been used in Wibral et al. (2013) to detect coupling delays in the bivariate case and assuming only one coupling delay. In Fig. 3.5(a) we show an example Venn diagram and time series graph to illustrate ITY. In analogy, we can also define a contemporaneous ITY

$$I^{\text{ITY}}_{X-Y} \equiv I(X_t; Y_t|\mathcal{P}_{Y_t}, \mathcal{N}_{Y_t}\backslash\{X_t\}, \mathcal{P}(\mathcal{N}_{Y_t}\backslash\{X_t\})). \tag{3.47}$$

ITY can be seen as a lag-specific transfer entropy (Wibral et al., 2013). In practical estimates of transfer entropy as in Eq. (3.39) actually mostly the vector is truncated at $\tau_{\max} = 1$ for which it corresponds to ITY at lag $\tau = 1$. ITY will be studied

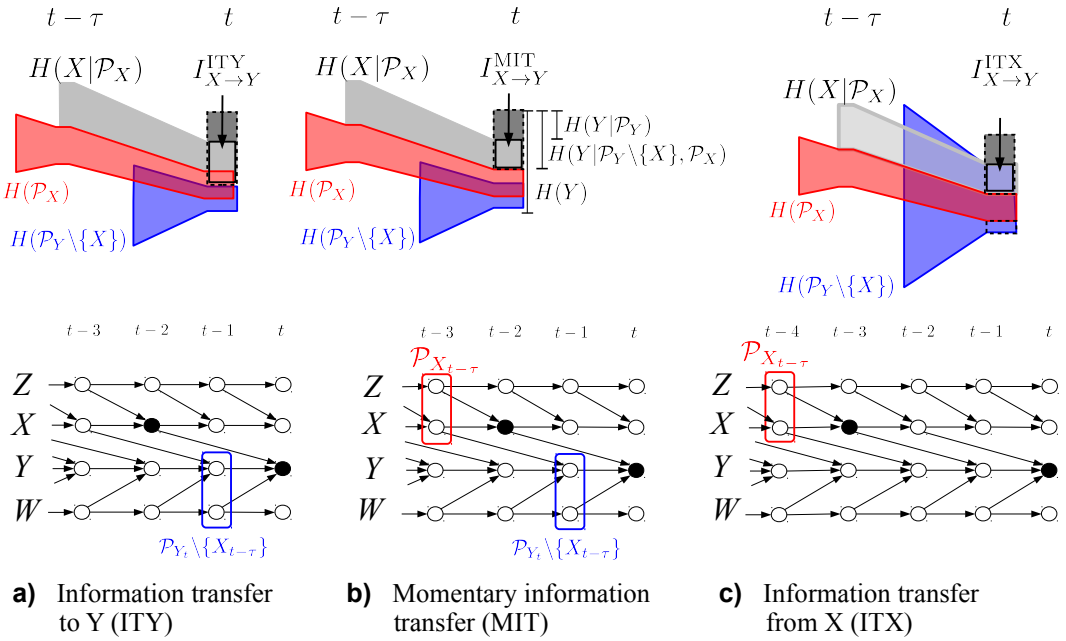


Figure 3.5.: Venn diagrams and time series graphs illustrating the measures ITY, MIT and ITX. The top panels show the entropy $H(Y)$ at time t (omitting t and τ in the labels) as a segmented column bar like in Fig. 3.2. It is composed of the source entropy $H(Y|\mathcal{P}_Y)$ (dark gray shaded) and parts of the source entropy $H(X|\mathcal{P}_X)$ (light gray shaded), the entropy $H(\mathcal{P}_X)$ of the parents of X (red), and the entropy $H(\mathcal{P}_Y\backslash\{X_{t-\tau}\})$ of the remaining parents of Y (blue), which may both overlap. In each panel, the respective CMI (solid framed segment) is the difference between the entropy in the dashed segment that includes transfer from X and the entropy that excludes it. For MIT in the middle this corresponds to the difference between the entropy $H(Y|\mathcal{P}_Y\backslash\{X\},\mathcal{P}_X)$ and the source entropy $H(Y|\mathcal{P}_Y)$. The lower panels show an example of a time series graph. In these graphs, the respective CMIs are between $X_{t-\tau}$ and Y_t (marked by the black dots) conditioned on either one or both of the parents $\mathcal{P}_{X_{t-\tau}}$ (red) and $\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}$ (blue). For the case of ITX (right panel (c)), we measure the information between two nodes that are not causally linked. Therefore all source entropy from X (light grey) enters $H(Y)$ only through its parents (blue entropy in Venn diagram).

in analytical and numerical examples in Sections 5.2.2 and 5.4 and its weakness for estimating conditional independence in the presence of autocorrelation will be discussed in Sect. 4.3. The ITY between Y and $X_{t-\tau} \in \mathcal{P}_{Y_t}$ can be related to the predictive power of $X_{t-\tau}$ compared to the other parents $\mathcal{P}_{Y_t}$. In Chapter 7, we study in how far the parents can be used as optimal predictors for time series forecasting.

3.4.5. Momentary information transfer

In this section, following Runge et al. (2012b), we introduce the concept of *momentary information transfer* (MIT) which is the main measure proposed in this thesis. The underlying concept of source entropy has been introduced in Section 3.1.3. MIT between X at some lagged time $t-\tau$ in the past and Y at time t is the CMI that measures the part of the entropy of Y that is shared with the source entropy of X relative to the entropy that excludes information from *both* parents:

$$\begin{aligned} I^{\text{MIT}}_{X\to Y}(\tau) &\equiv I(X_{t-\tau}; Y_t | \mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}) \\ &= H(Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}) - H(Y_t|\mathcal{P}_{Y_t}). \end{aligned} \tag{3.48}$$

This approach of "isolating source entropies" is sketched in a Venn diagram in Fig. 3.5(b). The attribute *momentary* (Pompe and Runge, 2011) is used because MIT measures the information of the "moment" $t-\tau$ in X that is transferred to Y_t. This "momentariness" is closely related to the property of coupling strength autonomy as we will show in Chapter 5. Similarly to the definition of contemporaneous links in Eq. (3.44), we can also define a contemporaneous MIT

$$I^{\text{MIT}}_{X-Y} \equiv I(X_t; Y_t | \mathcal{P}_{Y_t}, \mathcal{P}_{X_t}, \mathcal{N}_{X_t}\backslash\{Y_t\}, \mathcal{N}_{Y_t}\backslash\{X_t\}, \mathcal{P}(\mathcal{N}_{X_t}\backslash\{Y_t\}), \mathcal{P}(\mathcal{N}_{Y_t}\backslash\{X_t\})) \tag{3.49}$$

where $\mathcal{N}$ denotes the contemporaneous neighbors given by Eq. (2.11). Contrary to the lagged MIT, due to the Markov property the contemporaneous MIT is equivalent to the formula defining contemporaneous links Eq. (3.44), however with a finite dimension of the conditions.

Because MIT uses the parents $\mathcal{P}_{Y_t}$ as conditions, it also fulfills the property of lag-specific causality proposed in Sect. 3.1.2. Further, as shown in Sect. 5.3.2, because MIT also is conditioned on the parents $\mathcal{P}_{X_{t-\tau}}$, it fulfills the property of coupling strength autonomy. The very definition of MIT already leads to a low-dimensional estimation problem without arbitrary truncation parameters making MIT also be practically computable. In Sect. 4.3, we show how MIT can be used to obtain more reliable significance tests in conditional independence tests.

Each of the CMIs introduced in the preceding sections are intended to measure a different aspect of the coupling between X and Y. In the analytical analysis of simple models (Chapter 5) we will discuss the interpretability of the different measures in detail and give climatological interpretations comparing MI, ITY and MIT in the applications of Chapter 6. An attempt of a thermodynamical interpretation of MIT is discussed in Sect. 5.5.2.

Linear partial correlation MIT and regression MIT In the linear case, MIT using partial correlation instead of CMI should be understood as *momentary variance transfer*. Then, MIT is the cross correlation of the residuals after $X_{t-\tau}$ and Y_t have been regressed on both the parents of $X_{t-\tau}$ and Y_t (see Sect. 3.3). The contemporaneous MIT in the linear case of an autoregressive model is equivalent to the

partial correlation of the residuals after regressing each process on its parents as shown in Sect. 5.3.2. In analogy, for every variable $Y \in \mathbf{X}$, we define a (multivariate) MIT regression where the parents $\mathcal{P}_{Y_t}$ are taken as regressors $\mathbf{U}_Y^{\mathrm{MIT}}$. The residual's covariance and inverse covariance matrix can then be estimated from the regression residuals. This approach will be studied in Sect. 5.2.1.

3.4.6. Time-conditional variants

Just like time series graphs can be defined for subsets of the time axis, also conditional dependence measures can be defined in that way. For a general measure $I(X;Y|Z)$ each argument can be estimated for selected time samples only. For the set of all time indices $\mathcal{T}$ we denote the subsets of selected indices as $\mathcal{T}_X$, $\mathcal{T}_Y$, $\mathcal{T}_Z \subseteq \mathcal{T}$. Then the *time-conditional CMI* $I(X_{t-\tau};Y_t|Z_{t-\tau_Z})$ is estimated only from the time indices t fulfilling

$$t-\tau \in \mathcal{T}_X \ \wedge \ t \in \mathcal{T}_Y \ \wedge \ t-\tau_z \in \mathcal{T}_Z. \tag{3.50}$$

For more conditions Z, more time indices are added accordingly. Here $\mathcal{T}_X$, $\mathcal{T}_Y$, $\mathcal{T}_Z$ can also comprise the whole set $\mathcal{T}$. As an example used in the climate applications, consider the estimate of CMI for $\mathcal{T}_X = \mathcal{T}$, $\mathcal{T}_Z = \mathcal{T}$ and with $\mathcal{T}_Y$ comprising only the winter months in $\mathcal{T}$. Then the influence of X in any month on only the winter months in Y conditional on any month in Z is measured. This approach will be used in the climate applications in Chapter 6. The sets $\mathcal{T}_X$, $\mathcal{T}_Y$, $\mathcal{T}_Z$ can also be chosen by imposing conditions on the variables X, Y, Z. For example, one could study the transfer of information between X and Y if Z is above some threshold.

3.5. Quantifying interactions along paths and between multiple processes

In this section, part of the novel contributions of this thesis, we introduce measures to quantify interactions between multiple processes in the time series graph.

3.5.1. Quantifying information flow along paths

The previous measures LINK, ITY and MIT are all solely defined for causal links and are zero for a pair of processes that is not directly linked in the time series graph. But an indirect path such as drawn in Fig. 3.5(c) is still causal in the sense that the path connecting them is directed. The measure depicted in Fig. 3.5(c) is the *information transfer from X*, denoted by ITX and defined as

$$I^{\mathrm{ITX}}_{X\to Y}(\tau) \equiv I(X_{t-\tau};Y_t|\mathcal{P}_{X_{t-\tau}}). \tag{3.51}$$

It measures the part of source entropy in $X_{t-\tau}$ that reaches Y_t on any path and is, thus, not directly causal anymore, yet in many situations we might only be interested

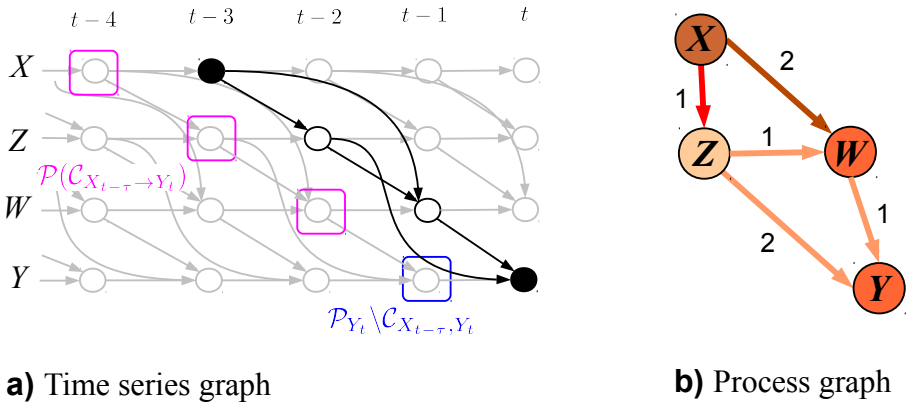


Figure 3.6.: (a) Time series graph and (b) process graph illustrating the momentary information transfer along paths (MITP). There are three causal directed paths connecting X_{t-3} and Y_t, two of length 2 via Z_{t-2} and W_{t-1} and one of length 3: $X_{t-3} \to Z_{t-2} \to W_{t-1} \to Y_t$. We denote the nodes along all directed or causal (see definition in Sect. 2.4.3) paths including $X_{t-\tau}$ by $\mathcal{C}_{X_{t-\tau}\to Y_t}$. The idea of momentariness then is to isolate all causal paths from the remaining process to assess the part of the source entropy of $X_{t-\tau}$ that is transferred on any causal path and shared with Y_t excluding those parents of Y that are not part of the causal path.

in the effect of X on Y, no matter how this influence is mediated. Note that in the Venn diagram of Fig. 3.5(c) it might seem, that ITX and MIT are the same, but the important difference is that the dashed entropies are not the same, i.e., ITX is measured relative to the larger entropy that includes information from $\mathcal{P}_Y$. In analogy, we can also define a contemporaneous ITX

$$I^{\text{ITX}}_{X-Y} \equiv I(X_t; Y_t | \mathcal{P}_{X_t}, \mathcal{N}_{X_t}\backslash\{Y_t\}, \mathcal{P}(\mathcal{N}_{X_t}\backslash\{Y_t\})). \tag{3.52}$$

ITX is studied analytically and numerically in Sections 5.2.2 and 5.4.

But ITX does not exclude information entering process Y_t from other sources. The idea of momentary information transfer was to isolate the information shared between two processes via a link from the remaining process. Now this idea can be generalized by isolating all causal (directed) paths from the remaining process to assess the part of the source entropy of $X_{t-\tau}$ that is transferred on any causal path and shared with Y_t, excluding those parents of Y that are not part of the causal path. Figure 3.6 illustrates this idea. With the nodes on all causal paths including $X_{t-\tau}$ (see Eq. (2.13) in Sect. 2.4.3) denoted by $\mathcal{C}_{X_{t-\tau}\to Y_t}$ the *momentary information transfer along paths* (MITP) is defined as

$$I^{\text{MITP}}_{X\to Y}(\tau) \equiv I(X_{t-\tau}; Y_t \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t}, \mathcal{P}(\mathcal{C}_{X_{t-\tau}\to Y_t})). \tag{3.53}$$

While there cannot be any directed paths between contemporaneous processes, one

can define a *contemporaneous MITP* where also intermediate processes on paths emanating from contemporaneous neighbors of $X_{t-\tau}$ to Y_t are included in the set $\mathcal{C}^*_{X_{t-\tau}\to Y_t}$. For example in Fig. 3.3, $\mathcal{C}^*_{Z_{t-1}\to X_t} = \{X_{t-1}\}$. There is no causal interaction between Z_{t-1} and X_t, but the entropy shared between Z_{t-1} and X_{t-1} due to their contemporaneous link is also shared with X_t, therefore the ITX $I(Z_{t-1}; X_t | Z_{t-2}, X_{t-2})$ is non-zero.

The analogous measure to ITY for paths is the *information transfer along paths* (ITP) defined as

$$I^{\text{ITP}}_{X\to Y}(\tau) \equiv I(X_{t-\tau}; Y_t \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t}). \tag{3.54}$$

ITP measures the transfer of *any* information entering $X_{t-\tau}$, conditioning out only those parents of the end node Y_t, that are not on any causal path. MITP will be studied on analytical examples in Sect. 5.2.3. In Sect. 5.5.4 we discuss how MITP can be used to quantify the influence of momentary perturbations entering the system at X on causally non-adjacent nodes in the time series graph. Climatological examples in Sect. 6.5 demonstrate the potential of this approach.

3.5.2. Quantifying interactions between multiple processes

Looking at Fig. 3.6, one immediate question is whether one can quantify how much of the information shared between X and Y went through Z and how much through W? Which of these is more important for explaining the indirect causal relationship between X and Y? The interaction information can be used to answer this question. For two processes $X_{t-\tau}$ and Y_t connected by a causal path, we define the *momentary interaction information* (MII) for $\tau > 0$ with an intermediate process Z for $\tau_Z > 0$ as

$$\begin{aligned}\mathcal{I}^{\text{MII}}_{X\to Y|Z}(\tau,\tau_Z) &\equiv \mathcal{I}(X_{t-\tau}; Y_t; Z_{t-\tau_Z} \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t}, \mathcal{P}(\mathcal{C}_{X_{t-\tau}\to Y_t})) &(3.55)\\ &= I^{\text{MITP}}_{X\to Y}(\tau) - \underbrace{I(X_{t-\tau}; Y_t \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t}, \mathcal{P}(\mathcal{C}_{X_{t-\tau}\to Y_t}), Z_{t-\tau_Z})}_{\text{MITP additionally conditioned on } Z_{t-\tau_Z}}\,. &(3.56)\end{aligned}$$

MII measures the effect of $Z_{t-\tau_Z}$ on the momentary information transfer along paths between $X_{t-\tau}$ and Y_t. Further versions of MII can be defined for different combinations of directed and contemporaneous links as shown in Appendix A.5.

In Section 5.2.4 we discuss several examples that demonstrate that MII is not necessarily always positive implying that an intermediate process can counteract the interaction between $X_{t-\tau}$ and Y_t. This measure can naturally be extended by including sets of processes from $\mathcal{C}_{X_{t-\tau}\to Y_t}$. Due to the symmetry of interaction information as defined in Eq. (3.28), also MII is symmetric in its arguments.

Again, we define the corresponding non-momentary measure to ITP named *interaction information along paths* (IIP)

$$\mathcal{I}^{\text{IIP}}_{X\to Y|Z}(\tau,\tau_Z) \equiv \mathcal{I}(X_{t-\tau}; Y_t; Z_{t-\tau_Z} \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t}) \tag{3.57}$$

$$= I^{\text{ITP}}_{X\to Y}(\tau) - \underbrace{I(X_{t-\tau}; Y_t \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t}, Z_{t-\tau_Z})}_{\text{ITP additionally conditioned on } Z_{t-\tau_Z}}. \tag{3.58}$$

Here the effect of Z on the transfer of *any* information entering in $X_{t-\tau}$ to Y_t is quantified. Climatological examples in Sect. 6.5 will show how MII and IIP can be used to quantify the importance of intermediate nodes in causal paths between two processes. In Sect. 5.5.4 we discuss how MII can be used as a measure of 'causal interaction betweenness', complementing concepts from complex network theory. In the linear case, we propose to use the absolute values of the partial correlation MITP and the second term in Eq. (3.55) and correspondingly for ITP to preserve the interpretation of positive and negative interaction information.

Decomposing ITY The information transfer to Y (ITY) and momentary information transfer are intimately related via interaction information. As shown in Sect. 5.3.2, MIT is always smaller equal than ITY and ITY can be decomposed by adding a "zero" in the following way

$$\begin{aligned} I^{\text{ITY}}_{X\to Y}(\tau) &= I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}) \\ &= \underbrace{I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}})}_{I^{\text{MIT}}_{X\to Y}} \\ &\quad + \underbrace{\big(I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}) - I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}})\big)}_{\mathcal{I}(X_{t-\tau};Y_t;\mathcal{P}_{X_{t-\tau}}|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\})}, \end{aligned} \tag{3.59}$$

where $\mathcal{I}(X_{t-\tau}; Y_t; \mathcal{P}_{X_{t-\tau}}|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\})$ is an interaction information that can be proven to be non-negative (via Theorem 5.2) and measures the enhancing influence of the parents of $X_{t-\tau}$ on the interaction of the causal link $X_{t-\tau} \to Y_t$.

3.5.3. Quantifying state-space based interactions

Another idea is to study the information not of X at a single lag, but at a joint vector of lags. A similar approach has also been proposed in Wibral et al. (2013) to measure the information in a *state* rather than a single observation that influences Y. Such a state could be X at $t-\tau$ together with its autodependency parents $\mathcal{P}^X_{X_{t-\tau}}$ reconstructed from applying the PC algorithm to the univariate process X similar to Ragwitz and Kantz (2002). We call this measure *state-information transfer to Y* (SITY)

$$I^{\text{SITY}}_{X\to Y}(\tau) \equiv I((X_{t-\tau}, \mathcal{P}^X_{X_{t-\tau}}); Y_t \mid \mathcal{P}_{Y_t}\backslash(X_{t-\tau}, \mathcal{P}^X_{X_{t-\tau}})). \tag{3.60}$$

3.6. Summary – the paradigm of conditional inference

In this chapter, we have discussed how causal interactions can be quantified based on the known time series graph. To this end, we have proposed a set of properties such a measure should fulfill and reviewed several known measures like lagged mutual information and transfer entropy and introduced novel measures, most importantly, the momentary information transfer. Table 3.1 recalls all introduced measures and lists their compatibility with the proposed properties as proven in Sect. 5.3. All information-theoretic measures fulfill the property of generality. The property of equitability is more a problem of estimation than of the theoretical measure and discussed in Sect. 4.2.5. Each measure is intended to quantify a different aspect of an interaction. The more conditional a measure, the more precise the underlying hypothesis and the better such a measure can be interpreted. The introduced measures provide a wide spectrum and can be viewed under the paradigm of conditional inference (Reid, 1995). One role of conditioning in statistics is the elimination of nuisance parameters in order to more accurately infer, for example, densities (Reid, 1995). In our case the measures implementing momentary information transfer are constructed to eliminate "external influences" as nuisance parameters and allow for more accurate measures of causal interaction strength. In Chapter 5, we will make this idea more precise with analytical examples and theorems. But before, in Chapter 4, we will discuss the important problem of estimation, where the properties of MIT will be shown to be advantageous to determine the statistical significance of causal links in conditional independence tests.

Table 3.1.: Summary of known and novel time series-based information-theoretic measures. If applicable, the second equation refers to the contemporaneous version. The last four columns provide an overview whether the measures fulfill the properties (3-6) discussed in Sect. 3.1.2. All information-theoretic measures fulfill the property (1) of generality. The property (2) of equitability is more a problem of estimation than of the theoretical measure and discussed in Sect. 4.2.5. In Chapter 5 we give counter examples and theorems proving that the measures satisfy or do not satisfy the axioms.

Measure class and name	Acro.	Definition	Conditional on	Properties (3–6) of interaction measures			
				lag specific	causal	coupl. str. autonomy	practically computable
MI of any two nodes							
Lagged mutual information	MI	Eq. (3.37)	none	Yes	No	No	Yes
CMI (Multivariate in the case of TE)							
Transfer entropy	TE	Eq. (3.38)	infinite past	No	Yes	No	No (truncated)
Decomposed transfer entropy	DTE	Eq. (3.42)	Markovian past	No	Yes	No	Yes
CMI of causal links							
Link-defining CMI	LINK	Eqns. (3.43, 3.44)	infinite past	Yes	Yes	No	No (truncated)
Information transfer to Y	ITY	Eqns. (3.45, 3.47)	Y's parents	Yes	Yes	No	Yes
State-information transfer to Y	SITY	Eqns. (3.60)	Y's parents	Yes	Yes	No	Yes
Momentary information transfer	MIT	Eqns. (3.48, 3.49)	*both* parents	Yes	Yes	Yes	Yes
CMI of causal paths							
Information transfer from X	ITX	Eqns. (3.51, 3.52)	X's parents	Yes	Yes	No	Yes
Information transfer along paths	ITP	Eqns. (3.54)	Y's parents	Yes	Yes	No	Yes
Momentary information transfer along paths	MITP	Eqns. (3.53)	parents along path	Yes	Yes	Yes	Yes
Interaction information of causal paths							
Interaction information along path	IIP	Eq. (3.57)	Y's parents	Yes	Yes	No	Yes
Momentary interaction information	MII	Eq. (3.55)	parents along path	Yes	Yes	Yes	Yes

Chapter 4.

Estimation

4.1. Introduction

However beautiful information theory with its generality is, estimating these measures from unfortunately finite data comes at a cost as will be explored in this chapter. In Section 4.2, we study the rather novel nearest-neighbor estimators of conditional mutual information (Frenzel and Pompe, 2007) which we extensively test regarding bias and variance. The interplay between these two impacts on the statistical power as a test for conditional independence to detect causal links. As summarized at the end of the chapter in Table 4.1, we find that – for the Gaussian model class studied here – CMI tests have good power up to as much as 32-dimensional conditions for a sample length of at least $1,000$. If we limit our scope to linear associations, independence tests with partial correlation have good power up to even higher dimensions (only studied up to 64 here) for as few as 100 samples. Further, we discuss the limits of estimating CMI with low bias as needed for an assessment of coupling strength. Here, we find that this more demanding task limits the CMI estimation dimension to about 8 for $1,000$ samples, while partial correlation is unbiased even for small sample lengths. In Section 4.3, we investigate analytical and shuffle tests for the significance of causal interactions for partial correlation and CMI, respectively, with an emphasis on the problem of autocorrelation which is ubiquitous not only in climate time series data (Von Storch and Zwiers, 2002) and commonly leads to an increase of false positives. Our main novel result here is that using momentary information transfer (MIT) to test for independence largely reduces this effect and entirely eliminates it for the linear partial correlation MIT. Also the assessment of confidence intervals is introduced. All these factors influence the causal inference algorithm explained in more detail in Section 4.4 where we demonstrate good detection rates in extensive numerical tests on a general class of nonlinear stochastic models. Finally, we discuss limitations, in particular the issue of multiple testing which we address by a two-fold significance test again utilizing momentary information transfer. Note that due to the efficient iterative testing scheme of the PC algorithm, the dimension of 32 mentioned above refers to the maximum number of *parents* in the causal graph, not to the number of processes which can be much higher.

Except for the results on numerical tests of the PC algorithm published in Runge et al. (2012a), the remaining sections consist of novel material.

4.2. Estimating conditional mutual information

4.2.1. Binning estimation

For symbolic data the choice of an estimator of entropy or (conditional) mutual information is straightforward by simply plugging in the symbol frequencies in the discrete versions of the formulas for entropies (Eq. 3.2) and conditional mutual information (Eq. 3.15). But the main focus of application here are variables taking a continuous range of values. Next to the class of "plug-in" estimators, in which the density is first estimated and then plugged into the entropy formula (Beirlant and Dudewicz, 1997; Hlaváčková-Schindler et al., 2007), a very popular method that allows to invoke discrete estimators is to quantize or partition the observation space into a set of bins (Paluš, 1996; Darbellay and Vajda, 1999; Steuer et al., 2002). In the simplest version, the joint and marginal distributions are estimated by binned histograms with some predefined partition (equidistant binning) into b bins. A more refined way is to use equiquantile bins where the bin edges are chosen such that the marginal distributions are uniform (Paluš, 1996). For the MI estimator proposed in Darbellay and Vajda (1999) the weak consistency was proven, i.e., MI can be approximated arbitrarily closely in probability.

However, binning estimators severely suffer from the curse of dimensionality because the number of joint bins B grows exponentially with the number of dimensions D and the number of bins b in the marginal dimensions as $B = b^D$, i.e., for $b = 3$, $D = 10 \Rightarrow B = 59049$. For common sample sizes of the order $\mathcal{O}(10^3)$, many bins are not populated enough resulting in heavily biased and usually overestimated values of MI (Hlaváčková-Schindler et al., 2007).

4.2.2. Nearest-neighbor estimation

Inspired by Dobrushin (1958), Kozachenko and Leonenko (1987) introduced a class of differential entropy estimators that can be generalized to estimators of conditional mutual information. This class is based on nearest neighbor statistics as further discussed in Kozachenko and Leonenko (1987); Frenzel and Pompe (2007). For a D_X-dimensional random variable X the nearest neighbor entropy estimate is defined as

$$\widehat{H}_X = \psi(T) + \frac{1}{T}\sum_{t=1}^{T}\left[-\psi(k_{X,t}) + \log(\epsilon_t^{D_X})\right] + \log(V_{D_X}) \tag{4.1}$$

with the digamma function as the logarithmic derivative of the gamma function $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$, sample length T, volume element V depending on the chosen metric, i.e., $V_{D_X} = 2^{D_X}$ for the maximum metric, $V_{D_X} = \pi^{D_X/2}/\Gamma(D_X/2+1)$ for euclidean metric with gamma function Γ. For every sample with index t, the integer $k_{X,t}$ is the number of points in the D_X-dimensional ball with radius ϵ_t. The formula holds for any ϵ_t and the corresponding $k_{X,t}$ which will be used in the following definition of a CMI estimator. Based on this entropy estimator, Kraskov et al. (2004)

derived two types of estimators for MI in the form of Eq. (3.10). One where the epsilon balls with radius ϵ_t are hypercubes (version V1) with side length $2\epsilon_t$ and one where they are hyper-rectangles (version V2) with a different side length in each dimension. The version V1 was generalized to an estimator for CMI first by Frenzel and Pompe (2007) and independently by Vejmelka and Paluš (2008). The CMI estimator is obtained by inserting the entropy estimator Eq. (4.1) for the different entropies in the definition of CMI in Eq. (3.15). For all entropy terms H_{XZ}, H_{YZ}, H_Z, H_{XYZ} in Eq. (3.15), we use the maximum norm and choose as the side length $2\epsilon_t$ of the hypercube the distance ϵ_t to the $k = k_{XYZ,t}$-nearest neighbor in the joint space $X \oplus Y \oplus Z$. The CMI estimate then is

$$\widehat{I}_{XY|Z} = \psi(k) + \frac{1}{T}\sum_{t=1}^{T}\left[\psi(k_{Z,t}) - \psi(k_{XZ,t}) - \psi(k_{YZ,t})\right]. \tag{4.2}$$

The only free parameter k is the number of nearest neighbors in the joint space of $X \oplus Y \oplus Z$ and $k_{xz,t}$, $k_{yz,t}$ and $k_{z,t}$ are computed as follows for every sample point indexed by t:

1. Determine (here in maximum norm) the distance ϵ_t to its k-th nearest neighbor (excluding the reference point which is not a neighbor of itself) in the joint space of $X \oplus Y \oplus Z$.

2. Count the number of points with distance strictly smaller than ϵ_t (including the reference point at t) in the subspace $X \oplus Z$ to get $k_{xz,t}$, in the subspace $Y \oplus Z$ to get $k_{yz,t}$, and in the subspace Z to get $k_{z,t}$.

Similar estimators, but for the more general class of Rényi entropies and divergences, were developed in Wang et al. (2009) and a conditional version in Schneider and Póczos (2012). To avoid points with equal distance, small amplitude random noise is added to break those ties.

This estimator uses the approximation that the densities are constant within the epsilon environment. Therefore, the estimator's bias will grow with k since larger k lead to larger ϵ-balls. The variance, on the other hand, becomes smaller for larger k because fluctuations in the ϵ-balls average out. The Kozachenko-Leonenko estimator is asymptotically unbiased and consistent (Kozachenko and Leonenko, 1987; Leonenko et al., 2008), a result that has been transferred to the MI estimator in Neureither (2013)[8]. Unfortunately, at present there are no results, neither exact nor asymptotically, on the distribution of the estimator as needed to derive analytical significance bounds (Hlaváčková-Schindler et al., 2007). In Goria and Leonenko (2005), some numerical experiments indicate that for many distributions of X, Y the asymptotic distribution of MI is Gaussian. But the important finite size dependence on the dimensions

[8] The previous works on Rényi entropy and divergence estimators in Kozachenko and Leonenko (1987); Goria and Leonenko (2005) contain an error that was corrected in Leonenko et al. (2008), but the (not straightforward) proof for the Shannon case was given in Neureither (2013).

D_X, D_Y, D_Z, the sample length T and the parameter k are unknown. In the next section we numerically evaluate bias and variance for the CMI estimator Eq. (4.2).

Some notes on the implementation: Before an analysis we standardize the time series, i.e., subtract their respective means and divide by the standard deviations in order to confine the probability densities. The theoretical CMI is invariant under this rescaling. In fact, it is invariant under any smooth and uniquely invertible transformation (isomorphism) as shown in Sect. 3.2.2. Kraskov et al. (2004), therefore, recommend to transform very skewed or rough distributions such that they become more uniform. Note, however, that the Kozachenko-Leonenko estimator and, hence, also the CMI estimator is unfortunately not invariant to more general monotonous transformations as needed in rank-transformations used in Pompe (1998). The main computational cost comes from searching nearest neighbors in the high dimensional subspaces which can be significantly speeded up using methods such as *kd-trie* (Vejmelka and Hlaváčková-Schindler, 2005).

4.2.3. Bias and variance

As mentioned in the introduction, the practical use of information theoretic quantities can be much limited if they cannot be reliably estimated. The MI estimator by Kraskov et al. (2004) has been quite extensively studied (Khan et al., 2007). One important property found is that it seems to be numerically unbiased even for finite samples of independent variables X, Y. This has been shown for a large range of values k and for many distributions and dimensions of X and Y. Some preliminary analyses in Frenzel and Pompe (2007) and Vejmelka and Paluš (2008) show, that the same holds for the CMI estimator, but only few parameters and low dimensions have been studied. Here we study a whole range of parameters, dimensions and sample lengths to explore the limits of CMI estimation. We use the following model:

$$\begin{aligned} X &= \frac{a}{\alpha}\sum_{i=1}^{D_Z} Z^{(i)} + \eta^X \\ Y &= cX + \frac{a}{\alpha}\sum_{i=1}^{D_Z} Z^{(i)} + \eta^Y, \end{aligned} \tag{4.3}$$

with Gaussian zero mean unit variances $\eta^{\cdot}$ and a multivariate D_Z-dimensional zero mean Gaussian process $\mathbf{Z}$ with covariance matrix $(\Sigma_{ij})_{1\le i,j\le D_Z}$ with $\Sigma_{i\neq j} = c_Z$, $\Sigma_{ii} = 1$, $i = 1, \ldots, D_Z$. The normalization $\alpha = \sqrt{D_Z + D_Z(D_Z-1)c_Z}$ is the standard deviation of $\mathbf{Z}$ and allows to compare the impact of different dimensions without changing the unconditional correlation (correspondingly for the Gaussian case also the mutual information) between X and Y, which is – independent of D_Z –

$$\rho(X;Y) = \frac{a^2(c+1)+c}{\sqrt{a^2+1}\sqrt{a^2(c+1)^2+c^2+1}}. \tag{4.4}$$

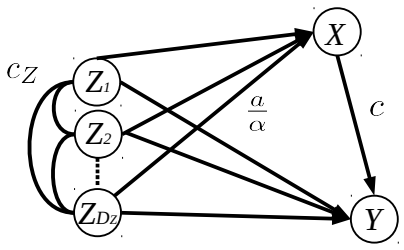
Figure 4.1(a) shows the graphical model for this coupling scheme. Note that the samples are i.i.d. for this model, in Sect. 4.3, we study the effect of autocorrelation in a time-dependent model. We varied the dimension D_Z and k with $T = 1000$ being fixed and conducted 1000 ensemble runs for every setup. The coefficient a determines the coupling strength of the driver $\mathbf{Z}$. We explore two different regimes, in Fig. 4.2 with $a = 0.5$ we study a weak common driving by $\mathbf{Z}$ and in Fig. 4.3 a strong driving with $a = 1$. In Fig. 4.1(b) we list the (unconditional) correlations for these parameter combinations which describe how confined the distribution is. The two columns correspond to the panels in Figures 4.2 and 4.3. In all cases, the common drivers are moderately correlated among each other with $c_Z = 0.3$.

Now we consider the CMI between X and Y conditional on $\mathbf{Z}$. The analytical CMI then is

$$I^{\text{theo}}(X; Y|Z^{(1)}, \ldots, Z^{(D_Z)}) = \frac{1}{2} \ln \left(1 + c^2\right). \tag{4.5}$$

For the top plots in Figs. 4.2 and 4.3, we set $c = 0$, therefore X and Y are conditionally independent and $I^{\text{theo}} = 0$ for all dimensions D_Z. The center and bottom plots show the relative bias $(\langle \widehat{I} \rangle - I^{\text{theo}})/I^{\text{theo}}$ for $c = 0.1$ ($I^{\text{theo}} = 0.005$) and $c = 0.4$ ($I^{\text{theo}} = 0.074$), respectively. The wire frames show absolute (top) and relative (center and bottom) 30% and 70% quantiles of the distribution.

For the weak forcing regime, we find that the estimator is almost unbiased for the conditional independent case (top panel in Fig. 4.2) while for stronger forcings (top panel in Fig. 4.3) the bias increases with D_Z and k. For very high dimensions (not shown here) it actually decreases again and becomes strongly negative. Therefore, unfortunately, the results found in Vejmelka and Paluš (2008) and Frenzel and Pompe (2007), that the estimator is unbiased if all processes are independent, do not apply for the conditional independent case and approximately zero bias seems to be only



a) Graphical model

coupling	weak driving $a = 0.5$	strong driving $a = 1$
$c = 0$ ($I^{\text{theo}} = 0$)	$\rho = 0.2$	$\rho = 0.5$
$c = 0.1$ ($I^{\text{theo}} = 0.005$)	$\rho = 0.29$	$\rho = 0.57$
$c = 0.4$ ($I^{\text{theo}} = 0.074$)	$\rho = 0.52$	$\rho = 0.72$

b) Model parameters

Figure 4.1.: (a) Graph of estimation model Eq. (4.3). (b) Table of model parameters and resulting correlation according to Eq. (4.4). The two columns correspond to the panels in Figures 4.2 and 4.3.

obtained if all variables $X, Y, \mathbf{Z}$ are at least only weakly correlated. The reason is that for strongly correlated variables the joint density is very confined leading to a bad approximation within the ϵ-balls.

For the dependent cases with $c = 0.1$ (center panels) and $c = 0.4$ (bottom panels) the same picture is obtained with moderate relative bias for weak forcings ranging between $+10\%$ for intermediate dimensions up to about -20% for $D_Z = 8$ and small k. For larger k the bias increases steadily. For strong forcings, the bias more strongly depends on k and D_Z, but is still moderate for small k for which the densities are better approximated. In Fig. 4.4, we show the relative root mean squared error, i.e., $\text{RMSE} = \sqrt{(\langle \widehat{I} \rangle - I^{\text{theo}})^2 + \langle (\widehat{I} - \langle \widehat{I} \rangle)^2 \rangle} / I^{\text{theo}}$ for $c = 0.4$ ($I^{\text{theo}} = 0.074$ to quantify the confidence in measuring the true CMI unbiased. Due to the decreasing variance the RMSE actually decreases for higher D_Z, but only for low k. Here an optimal value for $D_Z \approx 8$ around $k = 3$ appears, but slightly higher values around $k = 10$ are more robust for different dimensions.

Two findings are important for the two main research questions:

1. The variance generally decreases faster with larger k and D_Z (and stronger couplings c) than the bias with the parameter k implying that there is an optimal value for which a weak coupling can best be distinguished from the conditional independent case. This will be important for using CMI as a measure of detecting conditional independence – the first of our main research questions – and will be discussed in the next section.

2. To be able to reliably measure coupling strengths – our second research question – we conclude that small k should be used for which there is only a slight negative bias growing with D_Z and, hence, the dimension should be kept as low as possible.

We chose Gaussian noise since it is very common in climate data (Von Storch and Zwiers, 2002), but other distributions will have other dependencies and even for different covariances among the conditions $Z^{(i)}$ the dependencies change. We have tried to derive the analytical distribution for the Gaussian case, but computing the distributions of the different k_{XZ}, k_{YZ}, k_{XYZ}, k_Z turned out to be intractable or possible only for small k which is, however, not what we need for conditional independence tests. We, therefore, have to leave this difficult problem as an open challenge for future research.

4.2.4. Power as conditional independence test

Measuring conditional independence among continuous random variables presents a challenging problem that is far from being solved (Bouezmarni et al., 2009; Póczos et al., 2012; Póczos and Schneider, 2012). It underlies all causal detection methods from the framework of general (i.e., linear as well as nonlinear) Granger causality. For the linear case many procedures exist with good properties, see e.g., Bouezmarni et al. (2009). But for more general, *nonparametric* in statistical terms, conditional

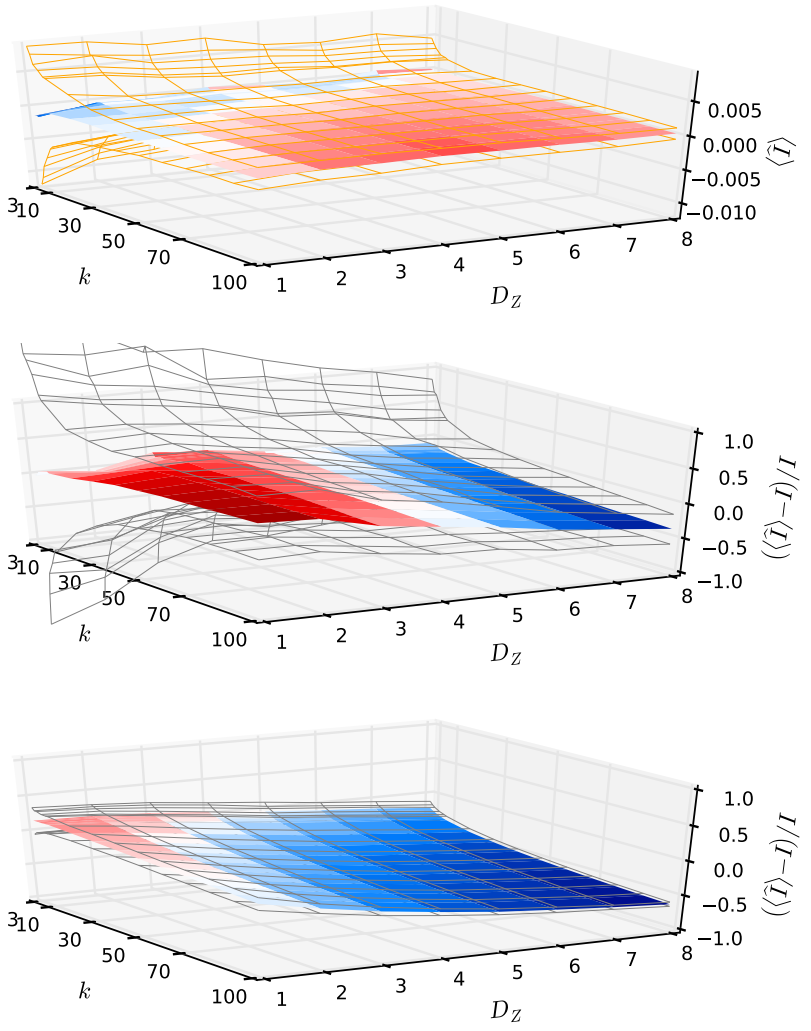


Figure 4.2.: Conditional mutual information estimation results for varying the nearest-neighbor parameter k and the condition dimension D_Z for a weak driving scheme with $a = 0.5$ in Eq. (4.3). The three panels correspond to the left column in Fig. 4.1(b). The top panel shows the absolute bias for $c = 0$ for which X and Y are conditionally independent and $I^{\text{theo}} = 0$ for all dimensions D_Z. The center and bottom panels show the relative bias $(\langle \widehat{I} \rangle - I^{\text{theo}})/I^{\text{theo}}$ for $c = 0.1$ ($I^{\text{theo}} = 0.005$) and $c = 0.4$ ($I^{\text{theo}} = 0.074$), respectively. The white shadings correspond to zero bias. The wire frames show absolute (top) and relative (center and bottom) 30% and 70% quantiles of the distribution.

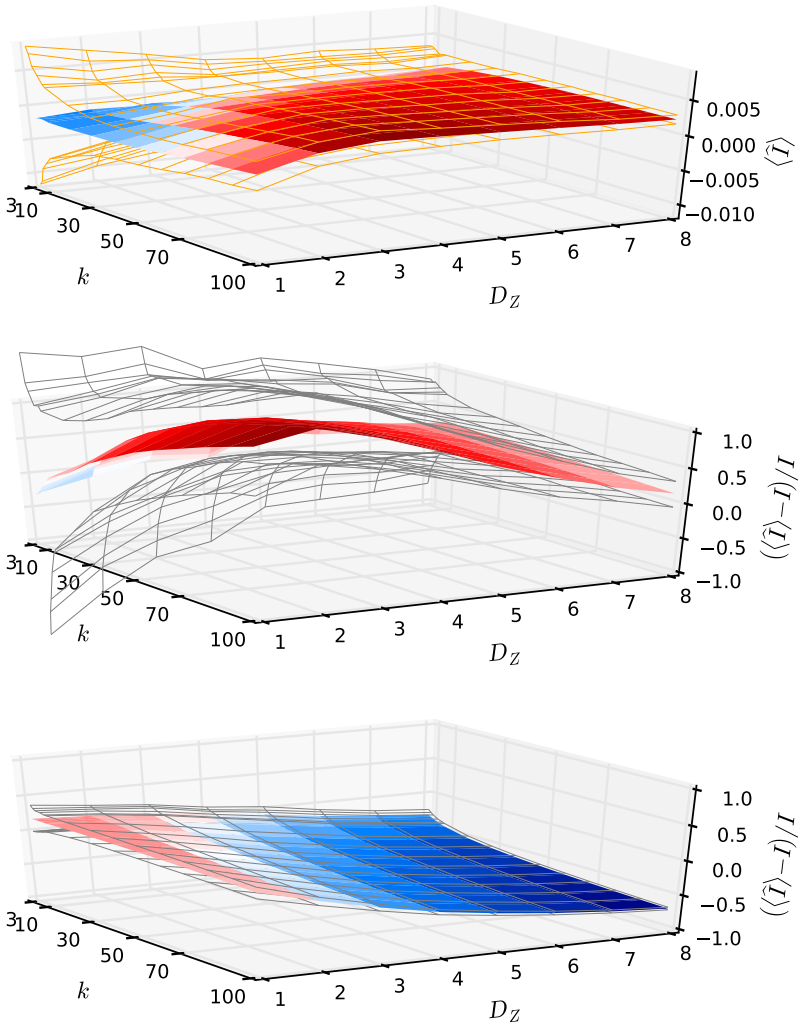


Figure 4.3.: As in Fig. 4.2, but for strong driving with $a = 1$ in Eq. (4.3). The three panels correspond to the right column in Fig. 4.1(b).

independence tests there are only few approaches (Bergsma, 2004; Bouezmarni et al., 2009). These tests are often based on the theory of *copulas* where the variables are transformed such that the marginal distributions are uniform (Bouezmarni et al., 2009; Póczos et al., 2012) and hence disentangle the dependency problem from the marginals. Recently, also Kernel based measures have been studied in the machine learning literature to test for conditional independence (Zhang et al., 2012).

As mentioned in the introduction, the problem of assessing conditional independence

has two separate aspects: (1) Building a statistical test under the null-hypothesis of conditional independence, which will be discussed in Sect. 4.3, and (2) having a measure with good *statistical power* which is the subject of this section. Statistical power, also known as sensitivity, in our case is the probability that the test will reject the null hypothesis of conditional independence if the two variables are actually conditionally dependent. Weak statistical power leads to a lot of false negatives (Type II errors), i.e., failing to detect a true causal link.

To this end, we investigate the *receiver operating characteristic* (ROC curve) shown in Fig. 4.5 which plots the sensitivity (probability of true positives – true link detection rate, 1− Type II error) against 1− specificity (probability of false positives – wrong link rate, Type I error). Every point on the ROC curve quantifies the overlap of the $c = 0.1$ distribution from the model ensemble described in the last section versus the conditional independent $c = 0$ distribution for a given threshold level and therefore measures what percentage of correct links one will get when allowing a certain rate of false positives. The higher this curve, the better.

The *area under the receiver operating characteristic* (AUC) provides a good aggregated measure of the statistical power (related to the Mann-Whitney test). Heuristically, AUC values above 0.8 are considered good and 0.5 implies no power at all, but different applications demand different preferences, i.e., either coverage (high detection rate) or high precision (low false positive rate as desired, e.g., in medicine). We will typically use significance levels that allow for 5% false positives. In Fig. 4.6, we show the two cases with weak (left column) and strong (right column) forcing. In the top row, we investigate the dependence of power on k and D_Z for sample length $T = 1000$. As expected from the sharp decrease of variance with k, we obtain good power for $k > 50$ after which the AUC is very robust and only slowly decreases for very large k. For increasing D_Z, on the other hand, the power steadily decreases

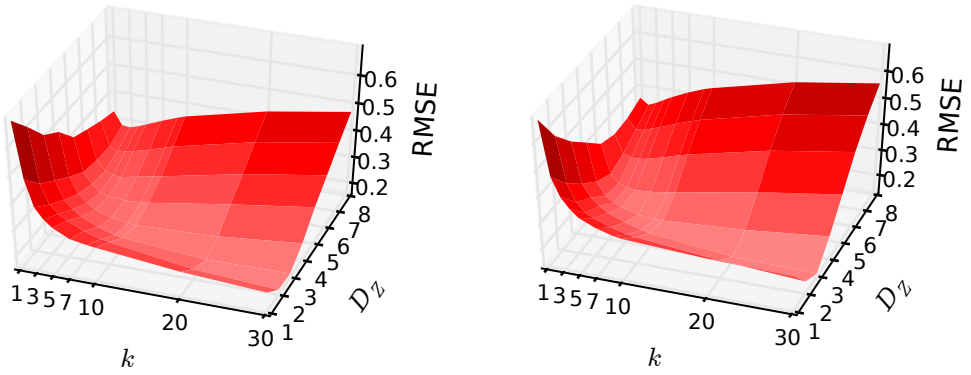


Figure 4.4.: Conditional mutual information estimation: Relative root mean squared error $\mathrm{RMSE} = \sqrt{(\langle \widehat{I} \rangle - I^{\mathrm{theo}})^2 + \langle (\widehat{I} - \langle \widehat{I} \rangle)^2 \rangle} / I^{\mathrm{theo}}$ for $T = 500$ and weak (left) and strong (right) forcing and with a coupling coefficient $c = 0.4$ ($I^{\mathrm{theo}} = 0.074$).

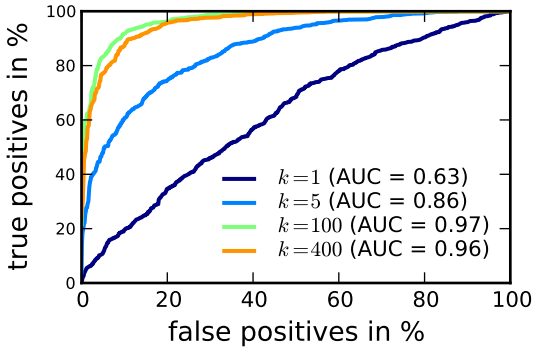


Figure 4.5.: Receiver operating characteristic (ROC curve) for $c = 0.1$ versus $c = 0$, $T = 1000$, and $D_Z = 6$ for the strong forcing regime. For small k, we observe very low power which stabilizes at high levels for $k > 100$.

(except for very small k), especially in the strong forcing regime. But even for a relatively large number of conditions $D_Z < 32$ we still have reasonable sensitivity. In the center row, we investigate whether the parameter k can be optimized with respect to the sample length T for fixed $D_Z = 6$. We observe an about symmetric behavior with optimal values at around $T/2$ but generally also values around $k = 50$ work well for all T (in Frenzel and Pompe (2007) values around $k/T = 0.02 \ldots 0.06$ are recommended). The power decreases strongly with smaller sample lengths below $T = 500$. In the bottom row, we fix $k = 50$ and explore the limits regarding sample lengths and dimensions D_Z. Good power in the "blue zone" is retained for $T > 500$ and $D_Z < 32$.

All these conclusions are to be interpreted with care since only a restricted type of models has been analyzed. We have tried to use a setup typically observed in climate data such that these findings can guide us in our choice of parameters in the analysis of real climate data. As mentioned before, finding convergence rates and exploring analytical properties of CMI nearest-neighbor estimators even for Gaussian densities is still an open problem in statistics.

Due to these problems, we pursue the analysis of causality and coupling strength also with linear partial correlation. Figure 4.7 demonstrates that even for small sample lengths partial correlation has good power up to very high dimensions. Note that we do not enter the regime where $D_Z > T$ here for which more sophisticated partial correlation estimation methods have to be utilized (Friedman et al., 2008).

4.2.5. Equitability and possible improvements

As mentioned in the introduction of Chapter 3, Reshef et al. (2011) put forward the heuristic property of equitability that a measure of dependence should fulfill. It implies that the measure should give similar values to equally noisy relationships. In Reshef et al. (2011), a large number of functional relationships $Y = f(X) + \eta$ is tested

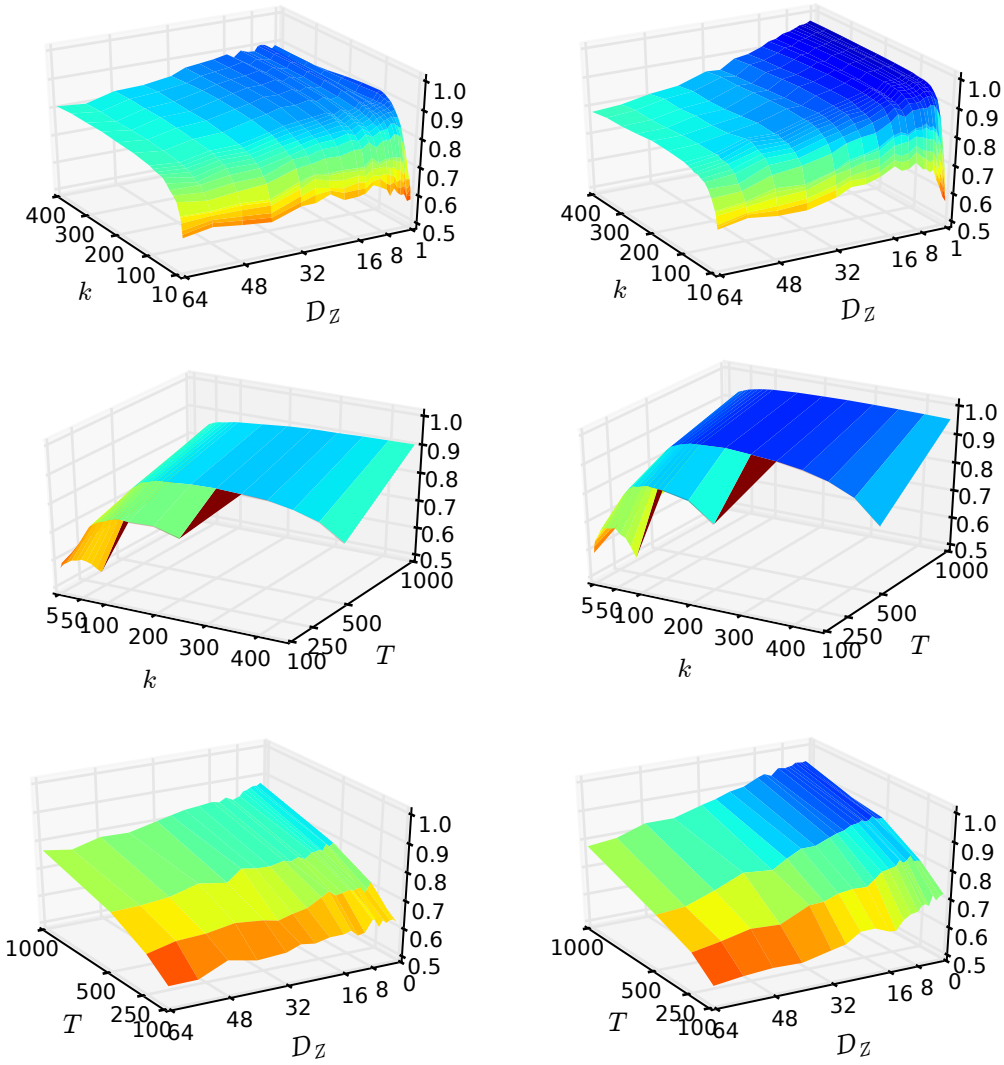


Figure 4.6.: Dependency of statistical power of the CMI estimator on k, D_Z, and T as measured by the area under the receiver operating characteristic (AUC) for $c = 0.1$ versus $c = 0$. The left column shows results for the weak forcing regime with $a = 0.5$ in Eq. (4.3) and the right column for strong forcing with $a = 1$. In the top panels $T = 1000$, in the center panels $D_Z = 6$ and in the lower panels $k = 50$. The color corresponds to the AUC value on the z-axis. Blue values above 0.8 mark parameter combinations with good power.

(for details see their paper) for which their *maximal information coefficient* (MIC) gives an almost linear decrease for increasing the standard deviation of the noise η. More precisely, they plot MIC versus $1 - R^2$, where R is the cross correlation between Y and the function $f(X)$. In their comparison with the mutual information estimator

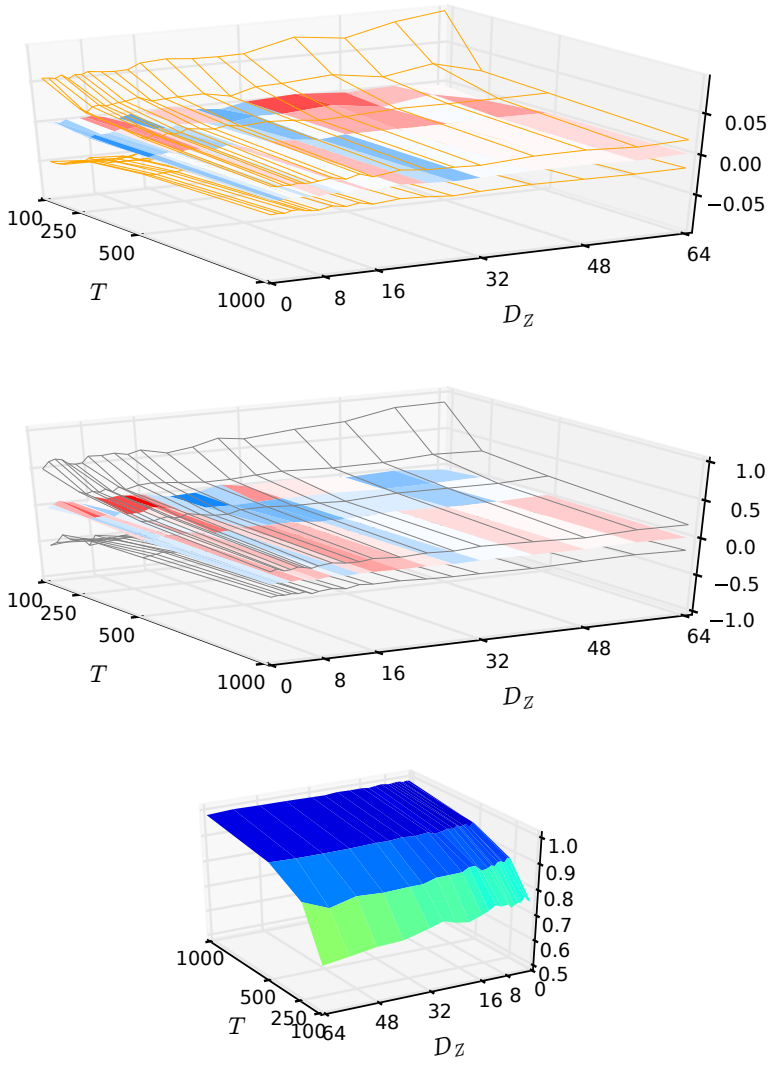


Figure 4.7.: (Top) absolute bias for $c = 0$ and (center) relative bias for $c = 0.1$ as in Fig. 4.2 and (bottom) statistical power as in Fig. 4.6 for partial correlation for the same model setup in the weak forcing regime.

of Kraskov et al. (2004), they find that MI does not scale linearly with the noise, but decays approximately exponentially. This implies that the same values of MI are given to low noise nonlinear and higher noise linear relationships, arguably an undesirable property shown in Fig. 4.8(a). In Fig. 4.8(b) we depict the rescaled MI according to the transformation $I \to \sqrt{1 - e^{-2I}} \in [0, 1]$ coming from the analytical relationship between I and the correlation ρ for Gaussian distributions in Eq. (3.36). Then MI scales quite linearly for most functional relationships. In our climatic examples we will use this rescaling to make the value of MI better comparable to correlations.

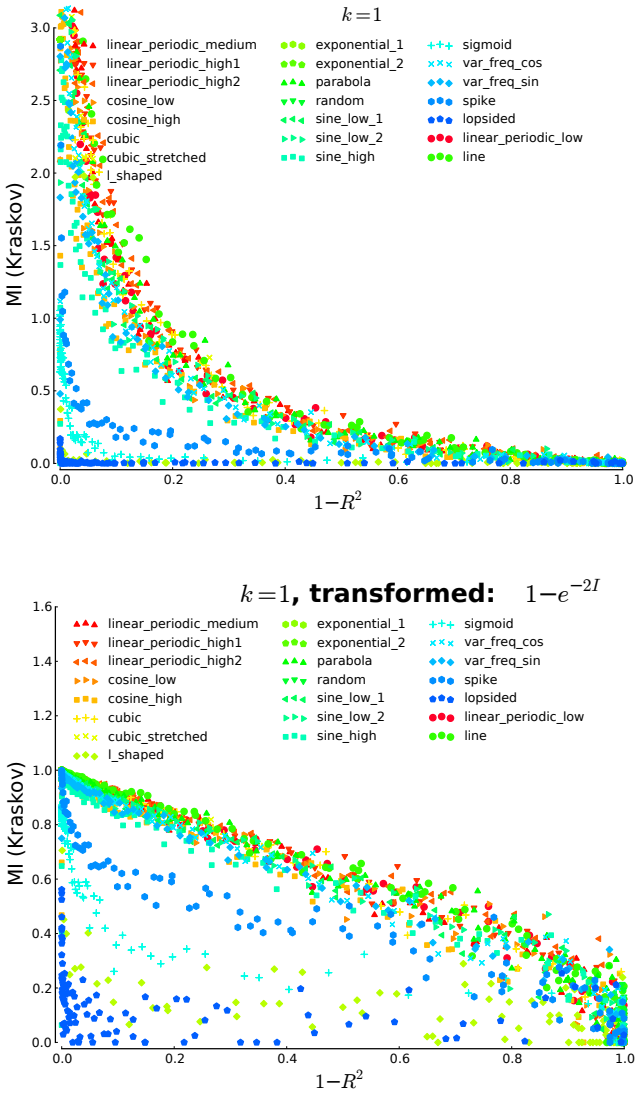


Figure 4.8.: The top panel shows the estimates of MI for $k = 1$ without rescaling and the lower panel the rescaled MI using the partial correlation transformation Eq. (3.36), which is squared here for better comparison to the R^2 value. For the model setup see Reshef et al. (2011). Essentially, Reshef *et al.* estimate the MI for a large number of functional relationships $Y = f(X) + \eta$ given in the legend and plot it versus the noise level given by $1 - R^2$, where R is the cross correlation between Y and $f(X)$.

Still, depending on the type of noise and for some pathological relationships shown in Fig. 4.8(b), for which also MIC gives similar deviations, the MI estimator looses the equitability property. We have undertaken many different attempts to improve the (C)MI estimator, from rescaling the distributions to generalizing the second estimator

version V2 from Kraskov et al. (2004) to the conditional case. This estimator uses hyper-rectangles instead of hypercubes which can approximate the local density better. But the analytical derivations of nearest neighbor statistics become very complicated in this case and we defer this challenging problem to future research.

4.3. Significance and confidence

4.3.1. Significance and autocorrelation

The statistical theory of significance testing is of great importance for building trust in hypotheses. Still a non-significant test does not bury a theory, often it just means that more information, more samples, are needed. But proper significance tests with well constructed null hypotheses allow to quantify the confidence into a certain result. One of the great advantages of linear theory is that an extensive framework for significance testing is available with ample analytical results. For information theory, few such theoretical results are available, especially not for the more recent advanced estimators as discussed in the previous section. Nevertheless, significance and confidence can be assessed using the framework of surrogate testing (Mudelsee, 2010).

Time series constitute a particular difficult case for significance testing because many analytical results assume that samples are *independent and identically distributed* (i.i.d.), which is often violated on the considered time scales in climate data due to autocorrelations, sometimes also termed serial correlation (see the example in Section 2.2.2). In many statistical methods the effect of autocorrelation is not desired and these methods are, therefore, modified to account for autocorrelation, for example, in the context of trend estimation (Zhang, 2004) and the detection of regime shifts (Rodionov, 2006) or change points (Wang, 2008). Also one usually accounts for autocorrelation in assessing the significance of a cross correlation (e.g., via permutation tests (Zwiers, 1990; Ebisuzaki, 1997)) because autocorrelation inflates the sample cross correlation coefficient even for independent time series. Further, it is known that for autocorrelated data the significance tests of adjacent lags in the cross correlation lag function are not independent anymore (Von Storch and Zwiers, 2002). Autocorrelation decreases the effective sample size and, thus, decreases the power of statistical tests as discussed in the previous Sect. 4.2.4. In the econometric literature, this problem is addressed using methods such as *first differencing*, where effectively the first derivative of a time series is used, or *pre-whitening*, where an autoregressive model is subtracted prior to further analyses. Since autocorrelation is ubiquitous in climate data, especially in time series of tropical temperatures, we investigate here how it affects the detection of statistically significant associations using MI, ITY and MIT. For ITY this is especially important since it is used in the PC algorithm to detect causal links, while for MIT we present novel results demonstrating how it can be used to obtain more reliable significance tests also under high autocorrelation. We study the problem of autocorrelation for linear as well as nonlinear sample estimators.

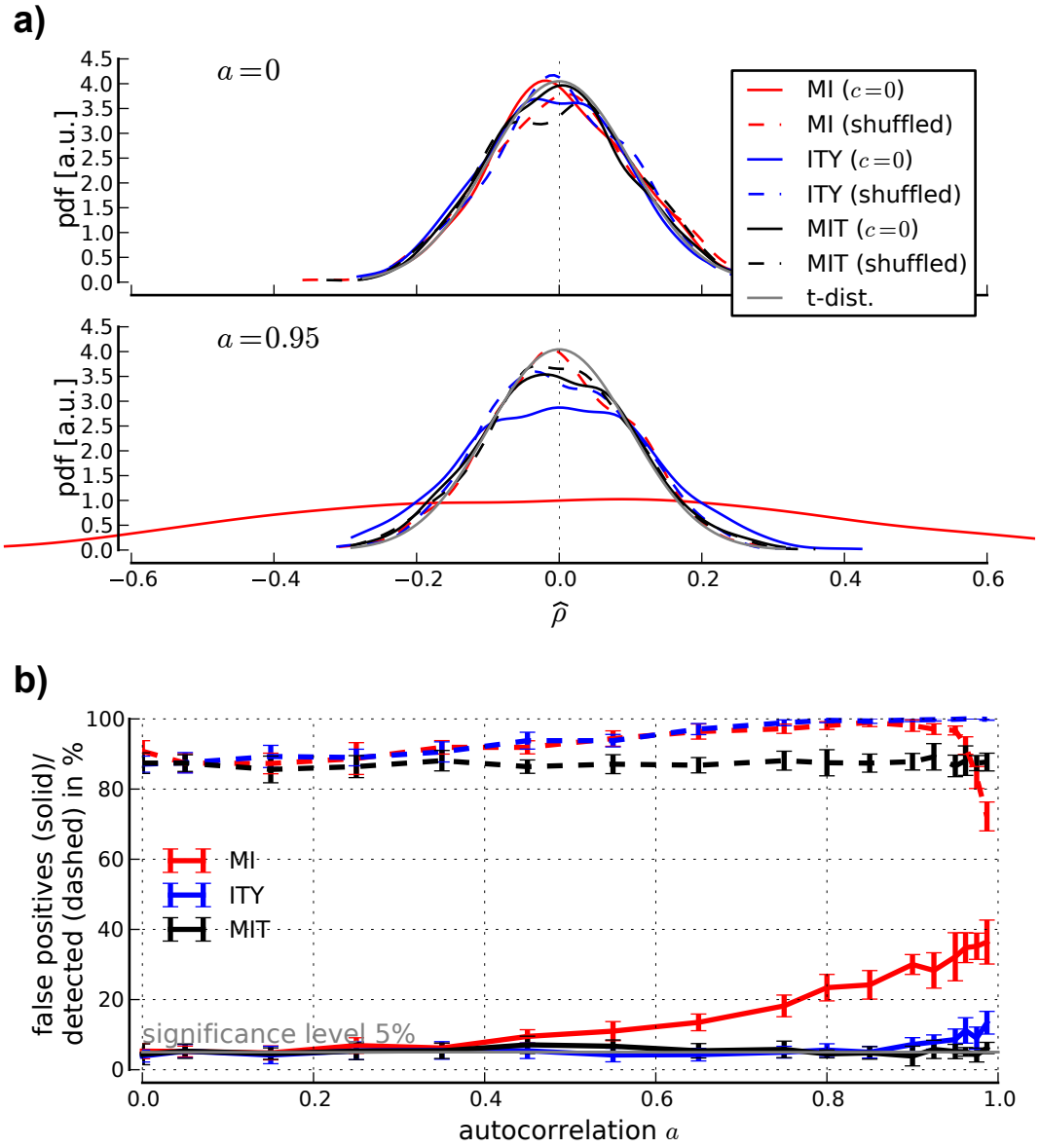


Figure 4.9.: (a) Distribution and (b) false positives and detection rate for (partial) correlation estimators for $T = 100$. (a) In the top panel we show Gaussian kernel density estimates of the distributions of $\widehat{\rho}$ for 1000 realizations of model (4.7) for $c = 0$ and zero (top) as well as strong autocorrelations (lower panel). (b) Solid lines denote the average actual false positives given a significance level $\alpha = 0.95$ (for $c = 0$) over 10 ensembles of 50 realizations, the error bar gives the standard deviation. The dashed lines give the true positive (detection) rate for a weak coupling $c = 0.3$.

4.3.2. Analytical partial correlation distribution

It is known, that the distribution of the partial correlation coefficient is the same as that of the cross correlation coefficient with the degrees of freedom reduced by the cardinality of the set of conditions q (Fisher, 1924). Therefore, the distribution of

$$\widehat{t}(YX|\mathbf{U}) = \widehat{\rho}(YX|\mathbf{U})\sqrt{\frac{n-2-q}{1-\widehat{\rho}(YX|\mathbf{U})^2}} \tag{4.6}$$

is Student's t with $n-2-q$ degrees of freedom with q being the dimension of $\mathbf{U}$. In the case of MIT, $q = |\{\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}\}|$. The assumptions underlying this result are Gaussianity and, importantly, i.i.d. samples.

To give a simple and very common example where this assumption is violated, consider the model Eq. (2.17),

$$\begin{aligned} X_t &= aX_{t-1} + \eta_t^X \\ Y_t &= aY_{t-1} + cX_{t-1} + \eta_t^Y , \end{aligned} \tag{4.7}$$

for $c = 0$ and strong autocorrelations $a = 0.9$ and where we assume the zero mean unit variance Gaussian innovations to be uncorrelated. The two processes are, therefore, independent, but the samples are serially dependent. We test the distribution of the cross correlation and the partial correlations ITY (where only the parents of Y are conditioned out) and MIT (where both Markov pasts are conditioned out). Here the parents are $\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\} = Y_{t-1}$ and $\mathcal{P}_{X_{t-\tau}} = X_{t-2}$ for $\tau = 1$. As shown in Fig. 4.9(a) for the cross correlation, this effectively reduces the degrees of freedom (Chatfield, 2013) and leads to an inflated sampling distribution. Also the distribution of the ITY $\rho(X_{t-2}; Y_t|Y_{t-1})$ is still inflated by autocorrelations in X because the residuals are *not* independent and the distribution is broadened due to less effective degrees of freedom. For larger sample sizes this effect becomes very small, though. Since the shuffle distributions for the cross and partial correlations are not inflated and match the analytical Student's t-distribution, this leads to an increased false positive rate as shown in the lower panel. The generation of shuffle surrogates is discussed in the next section.

On the other hand, the MIT estimator is not inflated by autocorrelation. This is also expected since the condition on the parents removes the dependency of X and Y on the past samples, the residuals $X_{\mathbf{U}}$ and $Y_{\mathbf{U}}$ given by Eq. (3.34) for a regression on both parents $\mathbf{U} = \{\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}\}$ are

$$\begin{aligned} X_{\mathbf{U},t} &= \eta_t^X \\ Y_{\mathbf{U},t} &= c\,\eta_{t-1}^X + \eta_t^Y , \end{aligned} \tag{4.8}$$

and, therefore, indeed serially independent since both η^X and η^Y are independent in time. This result can actually be generalized in the *coupling strength autonomy theorem* discussed in the next chapter and establishes that the MIT estimator in

many cases has the desirable property that the i.i.d. assumption is fulfilled in the sense that the samples are *conditionally i.i.d.* This has the important advantage that the analytical Student's t-distribution shown in Fig. 4.9(a) is valid and can be used for significance testing instead of computationally expensive surrogate tests which try to preserve serial dependency such as block-surrogates (Mudelsee, 2010).

In Fig. 4.9(b), we show that for increasing autocorrelation the false positive percentage for a (one-sided) significance level of $\alpha = 0.95$ is indeed 5% for the partial correlation MIT, while for ITY and especially MI more false positives occur than expected which makes the significance tests unreliable. The positive bias due to autocorrelation in MI and ITY actually increases the detection rate to some extent, albeit for very high autocorrelations it sharply decreases for MI. This might seem as an advantage of statistical power, but a desired property of a statistical test is that its power does not depend on external parameters. This reliability feature of MIT, independent of the strength of autocorrelation, can be used for independence tests in the PC algorithm since it allows for a more accurate significance test as discussed in Sect. 4.4.4.

4.3.3. Shuffle distribution for conditional mutual information

For the information-theoretic quantities no theoretical results about the sampling distribution exists. Here, we propose a shuffle test to at least approximate this distribution. For MIT, we shuffle (randomly permute) the samples of $\{X_{t-\tau}, \mathcal{P}_{X_{t-\tau}}\}$ against the samples $\{Y_t, \mathcal{P}_{Y_t}\}$. In this way the dependencies between X and its parents and Y and its parents are preserved. The dependencies between the parents, however, are also destroyed. For ITY used in the PC algorithm, we only shuffle $X_{t-\tau}$, also preserving the dependency between Y and its parents. In the supplement of Janzing et al. (2013) there is a proof that this procedure indeed makes the shuffled X independent of the unshuffled Y with its parents. The optimal approach would be to keep the conditions fixed and shuffle the pair $(X_{t-\tau}, Y_t)$ conditionally for each value of the conditions, but then one would have to use a histogram approach to assign values to the high-dimensional conditions. If a condition is parent to both X and Y, we randomly assign it to either shuffle set.

We generated 500 realizations of model (4.7) for $c = 0$ with a time series length $T = 1000$ and $k = 50$ and increasing a. For every realization, we used 100 samples in the shuffle test to estimate the 95% significance level. The results are shown in Fig. 4.10(a). In the top panel for zero autocorrelations the shuffle distributions of $\widehat{I}^{\mathrm{MIT}}$ and $\widehat{I}^{\mathrm{ITY}}$ well match the corresponding distributions under the true null hypothesis $c = 0$. Since the variance of the estimator decreases with the dimension (see Section 4.2.3), the distributions for $\widehat{I}^{\mathrm{MI}}$ are slightly broader (also seen in the climatological application Sect. 6.6). Thus, the false positive rate at a 95% significance level reliably yields about 5% false positives as shown in the lower panel. For stronger autocorrelations shown in the lower panel, however, the true $c = 0$ distribution of $\widehat{I}^{\mathrm{MI}}$ has a large bias and even larger variance (beyond the plotted range) and also for $\widehat{I}^{\mathrm{ITY}}$ and $\widehat{I}^{\mathrm{MIT}}$ we observe deviations from the shuffle distribution, but to a much

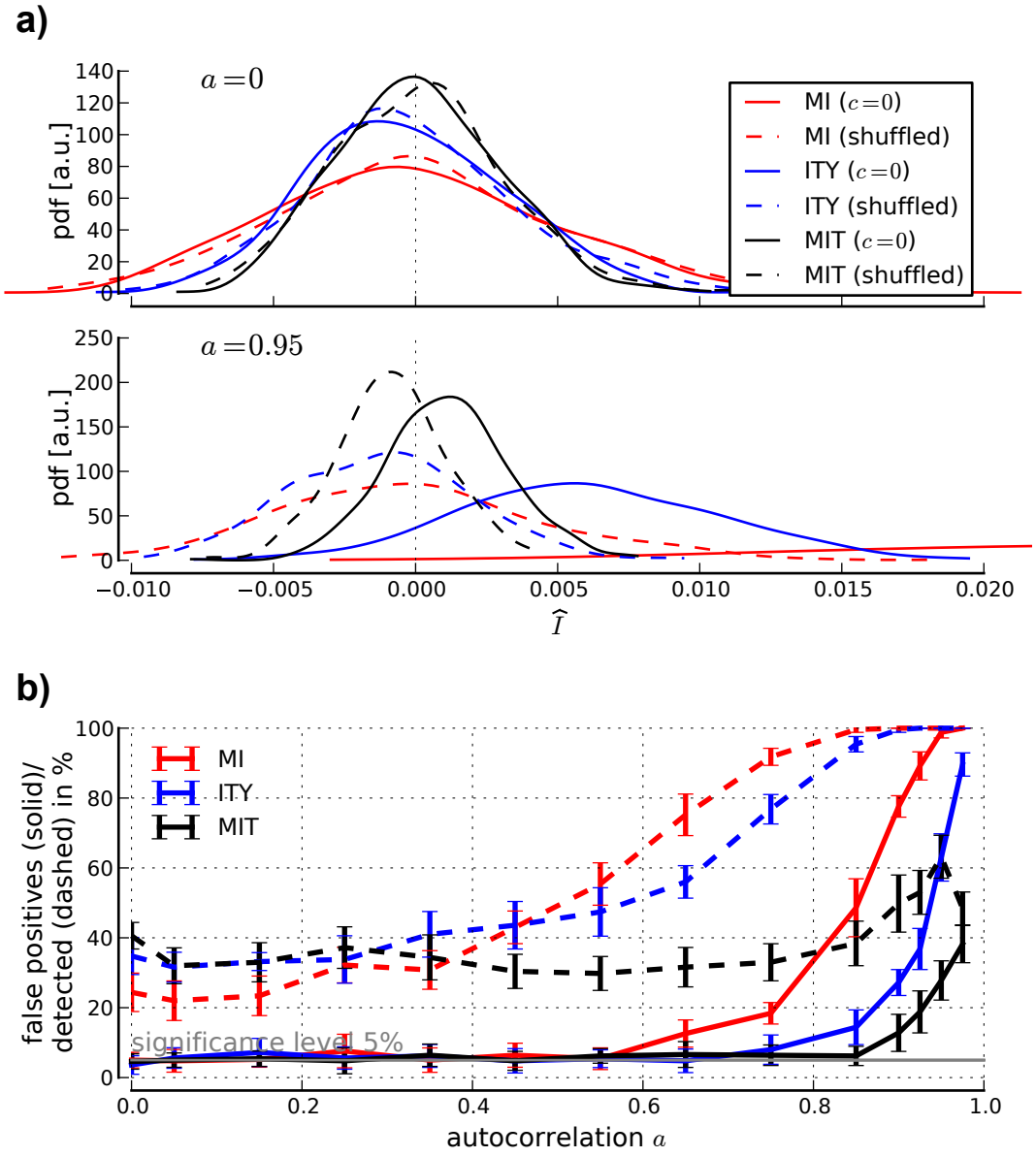


Figure 4.10.: (a) Distribution and (b) false positives and detection rate for MI and CMI estimators for $T = 1000$ and $k = 50$. (a) In the top panel we show Gaussian kernel density estimates of the distributions of $\widehat{I}$ for 1000 realizations of model (4.7) for $c = 0$ and zero (top) as well as strong autocorrelations (bottom). The distribution of MI (red) extents well beyond the plotted range. (b) Solid lines denote the average actual false positives given a significance level $\alpha = 0.95$ over 10 ensembles of 50 realizations, the error bar gives the standard deviation. The dashed lines give the average true positive (detection) rate for a weak coupling $c = 0.1$.

lesser extent in the case of MIT. Interestingly, all shuffle distributions have about the same small negative bias and differ only in their variance.

In Fig. 4.10(b), we show the false positive rate and percentage of detected couplings for $c = 0.1$ and increasing autocorrelation parameter a. In contrast to the partial correlation MIT, here also the false positive rate of MIT increases, but only for very large autocorrelations and much less than that of MI which is above 40% already for moderate autocorrelations of $a = 0.85$ and reaches almost 100%. Also the rate of ITY strongly increases earlier than that of MIT up to 80% while MIT maximally attains 40%. The high false positives for MI and ITY are a rather undesirable feature because, if a large number of pairs are tested, it leads to a bias towards strongly autocorrelated pairs even if all pairs are independent. This also affects the estimate of transfer entropy (Eq. (3.39)) which is commonly truncated at $\tau_{\max} = 1$ for which it is equivalent to the ITY studied here. For smaller k, MIT even better excludes the effect of autocorrelation because the smaller ϵ-balls better approximate the highly skewed distribution, but then the estimator's variance increases and the detection rate decreases. Also here, the bias leads to higher detection rates for MI and ITY and the rate, thus, depends on the coefficient a of the pair of processes which also gives a bias towards autocorrelated pairs. In our examples we choose $k = 100$ in the algorithm which is a balance between keeping the false positives at the desired significance level even in the presence of autocorrelation and keeping the variance low to preserve the detection rate.

4.3.4. Confidence bounds via bootstrapping

Apart from significance testing, in the climate analyses of Chapter 6 we also provide bootstrap confidence bounds (Efron and Tibshirani, 1993) which allow to quantify the uncertainty in the sample. These are computed from surrogates by drawing samples with replacement from the jointly lagged sample. For example, a surrogate for the partial correlation estimate $\widehat{\rho}(X_{t-1}; Y_t | Y_{t-1})$ for a sample length of T is created by randomly choosing T triples of lagged samples (X_{t-1}, Y_t, Y_{t-1}) and estimating their partial correlation. For conditional mutual information, however, this approach is not possible, because the multiple occurrence of the same sample point would yield tied values. Since tied values are broken by adding small random noise in the nearest-neighbor estimator, the surrogates would be biased towards lower values. Therefore, rather than drawing with replacement from the sample values (X_{t-1}, Y_t, Y_{t-1}), we draw from the nearest-neighbor counts $k_{xz,t}$, $k_{yz,t}$ and $k_{z,t}$ and conduct a bootstrap test only for estimating the mean in Eq. (4.2). This will at least approximate the uncertainty in estimating CMI to some extent. Confidence intervals will be used in several examples in the climate applications in Chapter 6.

4.4. Estimation of time series graphs

4.4.1. Practical implementation of causal algorithm

The PC algorithm to estimate time series graphs was already introduced in Section 2.4.6, here we discuss some more details and give a numerical example. In Section 2.4.6, we mentioned that "some measure" $I(X;Y|\mathbf{Z})$ can be used to assess the conditional independence necessary in the iterative algorithm. In the last chapter, we developed the information-theoretic apparatus providing such measures. But the PC algorithm can also be estimated in a linear version in combination with *partial correlation* to measure conditional linear dependence. In the climate applications we will pursue both approaches, partially because of the insights regarding estimation reliability in this chapter.

The free parameters of the modified PC algorithm for time series are the maximum lag $\tau_{\max}$, the initial number of conditions n_0, the significance threshold I^* to determine whether $I(X_{t-\tau};Y_t\,|\,\tilde{\mathcal{P}}_{Y_t}^{n,i}) > 0$ (and correspondingly for contemporaneous links), and the k-nearest neighbor parameter of the CMI estimator. To speed up the performance in the i-loop we propose to first test links with weakest CMI conditioned on nodes with largest CMI. Weakest and largest CMI are determined after each n-loop by sorting the elements $X_{t-\tau}$ in $\tilde{\mathcal{P}}_{Y_t}$ by the value of $\min_i |I(X_{t-\tau};Y_t\,|\,\tilde{\mathcal{P}}_{Y_t}^{n,i})|$ (analogously for $\tilde{\mathcal{N}}_{Y_t}$ in the algorithm for the contemporaneous graph). $\tau_{\max}$ can be chosen very large so as to include all possible coupling delays, since it will not increase the estimation dimension in the algorithm. In the original algorithm, the initial n_0 is set to one, which has a high probability to be too small to unveil spurious links, leading to a bad performance because the algorithm runs through all possible conditions without reducing the set of parents. Ideally, n_0 is initially chosen as large as the expected number of parents. This will much faster eliminate spurious links and also alleviate the problem of multiple testing as we found in numerical experiments. The significance threshold can either be subjectively set in every iteration or be computed from a significance test as discussed in the last section. Note that this test is separate for every lag τ since the conditions can vary, for example for the test of contemporaneous links. Further parameters to limit computational time especially for large networks are the maximum number of conditions $n_{\max}$ and $n_i = 5$, the number of tests per n.

4.4.2. Example

We demonstrate the method with a system of four stochastic delay-differential equations and couple them linearly and nonlinearly, also in the stochastic terms. This system of Ornstein-Uhlenbeck processes can be interpreted as nonlinearly coupled particles, each fluctuating in its harmonic potential:

$$\dot{X} = -0.5\ X(t) + 0.6\ W(t-4)\cdot\eta_X(t)$$
$$\dot{Y} = -0.9\ Y(t) - 1.0\ X(t-2) + 0.6\ Z(t-5) + \eta_Y(t)$$
$$\dot{Z} = -0.7\ Z(t) - 0.5\ Y(t-6) + \eta_Z(t)$$

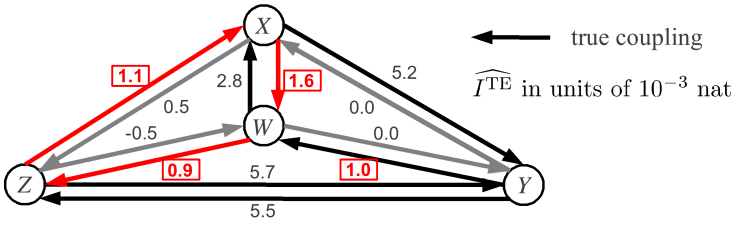


Figure 4.11.: Estimated TE between all subprocesses X, Y, Z, W for an example process Eq. 4.9 with true coupling structure given by the black arrows. Of the four estimates of comparable size marked in red, only the link $Y \to W$ is a correctly identified coupling.

$$\dot{W} = -0.8\ W(t) - 0.4\ Y(t-3)^2 + 0.05\ Y(t-3) + \eta_W(t) \tag{4.9}$$

with independent unit variance white noise processes $\eta.(t)$. Thus, we have a bidirectional feedback $Y \rightleftharpoons Z$ and a feedback loop $X \to Y \to W \to X$ in which $Y \to W$ is nonlinear and a stochastic coupling $W \to X$. This system would be hard to analyze using model-based approaches, especially given short sample lengths as used here ($T = 1000$). Throughout the analysis, we have used a fixed significance threshold $I^* = 0.015$ and a maximum lag $\tau_{\max} = 15$. The shuffle test will be extensively tested in the next section.

First, we demonstrate that transfer entropy (TE) in its common definition Eq. (3.39) suffers from the "curse of dimensionality" which strongly affects the reliability of causal inference as demonstrated in Fig. 4.11. There we estimated TE between all subprocesses X, Y, Z, W with true coupling structure given by the black arrows. If we truncate the infinite past vectors used in the common definition of TE at $\tau_{\max} = 15$ lags, the estimation dimension is 61. The true TE of all gray and red links is zero, since they do not represent direct couplings. So of the four estimates of comparable size marked in red, only the link $Y \to W$ is a correctly identified coupling. TE is further tested numerically in Sect. 5.4.

In the PC algorithm, on the other hand, the estimation dimension is only iteratively increased. Fig. 4.12 shows the iteration steps. Guided by our investigations on statistical power, for the present sample size $T = 1000$ we use the CMI estimation parameter $k = 100$. Step (0.0) gives the result of an analysis using only MI, the first step of the algorithm. We would wrongly infer that Y drives X, X drives W, X and Z are coupled and conclude on a long-range memory process within Y and Z at $\tau \approx 12$. Also the precise coupling delays are buried under a broad range of significant lags similar to what we have seen in the climatological example in Section 2.2.2. The algorithm proceeds as follows. Considering the estimation of the parents of X, we choose an initial $n_0 = 3$ and start with the links with weakest MI, Z at $\tau = 1, 2$, conditioned on the three preliminary parents with largest MI, $\tilde{\mathcal{P}}^{3,0}_{X_t} = \{X_{t-1}, X_{t-2}, W_{t-4}\}$. As these links are due to Y which drives Z and with one step delay also X via W (see process graphs in Fig. 4.12), $I(Z_{t-\tau}; X_t \,|\, \tilde{\mathcal{P}}^{3,0}_{X_t})$ for

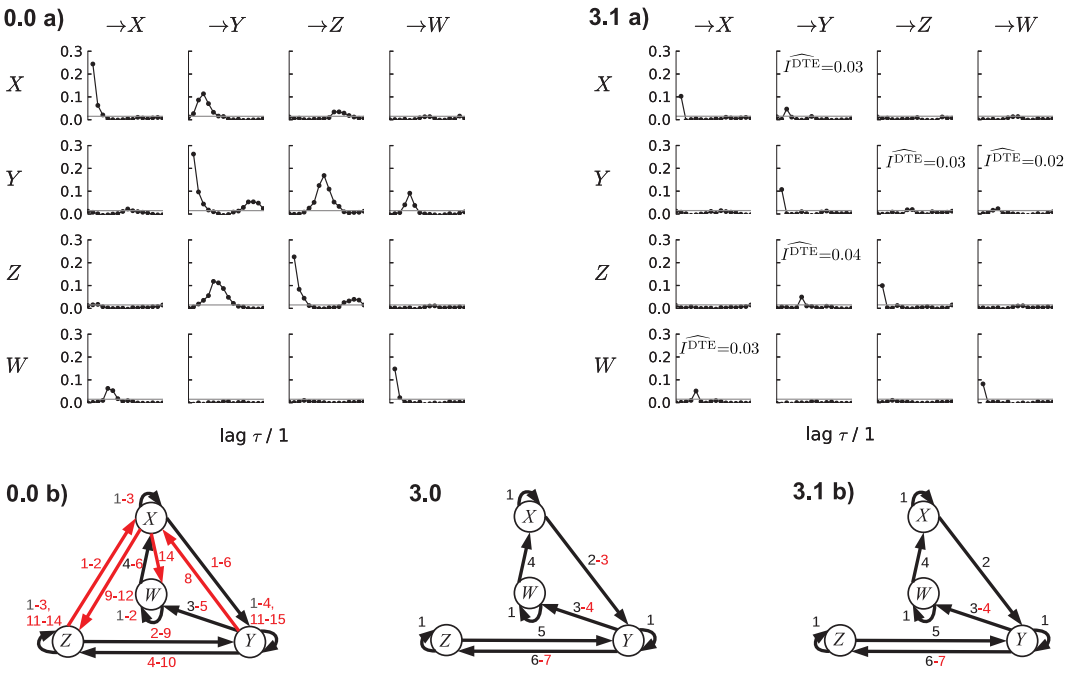


Figure 4.12.: Iterative steps in the analysis of model Eq. (4.9), time series length $T = 1000$, integration time step $dt = 0.01$, sampling interval $\Delta s = 100$, MI and CMI estimated using $k = 100$. The label $(n.i)$ indicates the iteration step, (0.0 a) shows the MI lag functions and (0.0 b) the process graph. With an initial $n_0 = 3$ the next step (3.0) with three conditions is already almost identical to the converged graph in step (3.1), where the values in the boxes denote the estimate $\widehat{I^{\mathrm{DTE}}}$ of TE via Eq. (3.42) with τ^* chosen such that $I(X_{t-\tau^*-1}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau^*-1}})$ has declined below significance. Incorrect links and lags are in red.

$\tau = 1, 2$ vanishes and we remove these links from $\tilde{\mathcal{P}}_{X_t}$. The second weakest link, $Y_{t-8} \to X_t$, is indirectly mediated via W and thus $I(Y_{t-8}; X_t \,|\, \tilde{\mathcal{P}}^{3,0}_{X_t})$ vanishes and we remove also this link. Next, we check the coupling from W and the auto-dependency from X_{t-2} conditioned on the same nodes, whereupon the links from W_{t-5} and X_{t-2} vanish. Now the set $\tilde{\mathcal{P}}_{X_t}$ is smaller than $n = 3$, i.e., all possible conditions have been tested, and the algorithm for X converges already in the second step. For Y, the coupling structure in step (0.0) is true, but with inaccurate lags and after the iteration converges in step (3.1), all peaks are sharp at the correct lag (see second column in (3.1 a)). The only difference between step (3.0) and (3.1) lies in the incorrect link $X_{t-3} \to Y_t$ which is due to the different conditions used: For step 3.0 we estimated $I(X_{t-3}; Y_t \,|\, Y_{t-1}, Z_{t-5}, Z_{t-6})$ (using the conditions with largest MI towards Y), while for step 3.1 we estimated $I(X_{t-3}; Y_t \,|\, Y_{t-1}, Z_{t-5}, X_{t-2})$ which reveals the link as indirect. The iterations for Z and W converge in the second step, again without the need to increase n.

Apart from some inaccuracy in the coupling lags, which is due to the continuous

nature of the system, this yields the correct graph. With this time series graph, we can also estimate DTE according to Eq. (3.42). While the dimension for the direct naive estimation of TE via Eq. (3.39) is $D = \tau_{\max} \cdot 4 + 1 = 61$, the dimension of DTE lies between 5 and 24 (depending on $\mathcal{S}_{Y_t, X_{t-\tau}}$). It is interesting to compare DTE with the model parameters in Eq. (4.9): $Z \to Y$ has a higher $\widehat{I^{\mathrm{DTE}}}$ than $X \to Y$, while the corresponding parameters are 0.6 and 1.0 respectively. TE as a measure of the total influence between two processes therefore cannot be simply related to the parameters of the underlying model. In Chapter 5, we will demonstrate that MIT can better be related to the model's coefficients and, thus, gives a better estimate of the strength of a mechanism.

4.4.3. Numerical experiments – detection and false positive rate

We investigated the performance of the algorithm also on a whole general class of model systems: discrete time, nonlinear stochastic delay models, also called generalized additive models in the statistics literature (Hastie and Tibshirani, 1986), here defined as follows:

$$X_i(t) = c_i\, X_i(t-1) + \sum_k c_k\, f_k^i(X_1(t-\tau_{1i}), \ldots) + \eta_i(t) \qquad i \in \{1,2,3,4\} \qquad (4.10)$$

with i.i.d. Gaussian white noise η_i. The f_k^i are products of $X_\cdot^p$ raised to a power $p \in \{0,1,2\}$. Therefore, linear terms of the form $f_k^i = X_\cdot$ and nonlinear terms like $f_k^i = X_{j_1}^2, \quad X_{j_1} \cdot X_{j_2}, \quad X_{j_1} \cdot X_{j_2}^2, \ldots$ for $j_i \in \{1,2,3,4\}$ are allowed. The possible lags are $\tau_{ji} \in \{1,2,3\}$ and the auto-dependency and coupling coefficients are $c_i, c_k = 0.1 \ldots 0.5$ with equal probability. With this setup, we ran 1000 numerical experiments, each with time series length $T = 1000$. For the analysis we used the CMI estimation parameter $k = 100$, $\tau_{\max} = 6$ and the shuffle test described in the last section calculated from 100 samples with $\alpha = 95\%$ as for CMI this is computationally already quite expensive.

In Fig. 4.13 the analysis of convergence in (a) shows that, as expected, setting an initial $n_0 = 3$ speeds up the performance compared to starting with $n_0 = 1$. More than 40% converge already in the second step. (b) implies that while the $n_0 = 0$ variant needs more iterations and thus multiple tests, slightly fewer tests are with higher dimensional conditions. As can be seen in (c), for MI we found – as expected – very large spurious false positive rates. The detection rates are quickly reaching 100% for strong enough links. The rates for MI are always higher. Since only those links with significant MI are further tested, the causal detection rates could also be seen relative to the non-causal rates.

4.4.4. Limitations

One concern with the PC algorithm is the problem of sequential testing. That is, if each link is tested multiple times at the same α-level, the resulting combined significance level is higher, which should be kept in mind when interpreting the false

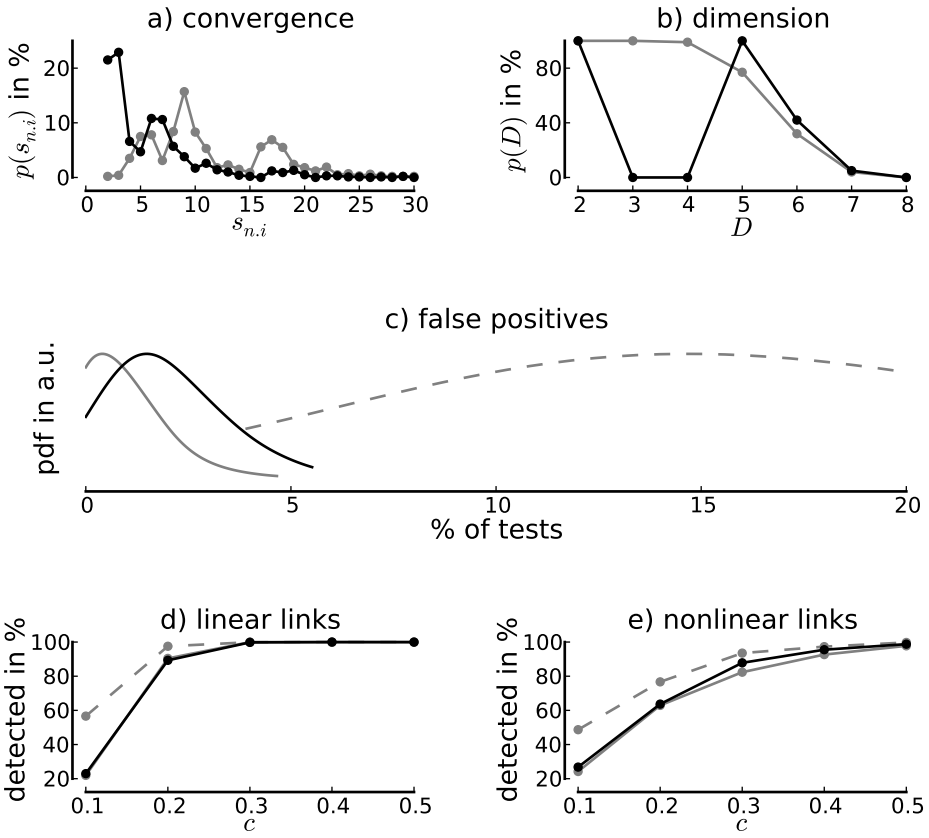


Figure 4.13.: Results of the numerical experiments. (a),(b) Convergence of PC algorithm for 1000 experiments. Gray solid lines denote the algorithm run with initial $n_0 = 1$ and black solid lines with $n_0 = 3$. (a) Histogram of the total number of iterations $s_{n.i}$ over the n- and i-loop until convergence. (b) Histogram of estimation dimensions $D = 2 + n$. The remaining plots summarize the detection performance by comparing the converged graph with the true graph. We also computed the unconditioned MI (gray dashed lines) for comparison. (c) Kernel density histogram (estimated using Scott's rule) of the percentage of false positives (as measured in % of tests) in each experiment. (d),(e) Percentages of correct detections for each coupling (or auto-dependency) strength parameter c for linear and nonlinear couplings.

positive rate of inferred links. This problem is usually treated by correcting for the number of tests (e.g., using a Bonferroni correction), but this number is not known a priori in the algorithm. We found that the $n_0 = 1$ variant with more iterations has a false positive rate far below the expected 5%, and therefore also misses out more correct links. However, in our experiment we found this to be true only for

some nonlinear links as demonstrated in Fig. 4.13(e). Note that the number of false positives also strongly depends on the number of processes considered (four in our simulations) and on $\tau_{\max}$.

To overcome this problem, we have used in some climate applications in Sect. 6.5 a two-fold significance evaluation. In the first step we use the PC algorithm to estimate the parents of each process, i.e., as a variable selection method. Then we use MIT in a second step and test all possible links again (not only the ones inferred by the PC algorithm using ITY). That way, we utilize the advantage of MIT regarding autocorrelations discussed in Sect. 4.3.1 (which can also come from other parents) *and* we only do one test per pair and can trust the significance level more, especially for partial correlation.

A further limitation are spurious causalities due to imperfect observations of states of the driving system (Smirnov, 2013). That is, even if Z is a common driver of conditionally independent X and Y, the CMI

$$I(X;Y|Z+\epsilon) \tag{4.11}$$

might not be zero because the noise ϵ deteriorates the full knowledge about Z needed to unveil it as a common driver.

4.5. Summary

In this chapter, we have found that the model-free generality of information theory comes at a cost. Even though superior to binning estimators, the recently introduced nearest-neighbor CMI estimators still demand much longer time series lengths compared to model-based measures and the number of possible processes that can be tested is limited. Further, the longer computational time hampers the work flow.

But there are also some good news as summarized in Tab. 4.1. We found that, despite the bias for weakly forced processes, statistical tests for conditional independence have quite good power even for rather high dimensional settings up to 32. Further, the problem of serial dependencies commonly occurring in climate data is drastically reduced using MIT while estimates of MI and ITY suffer much more from high autocorrelation. Also our numerical experiments demonstrate high detection rates even for nonlinear dependencies. The two-step significance procedure discussed in the previous Sect. 4.4.4 alleviates the problem of multiple testing and allows for a higher confidence in significance levels. Regarding the question of how many processes can maximally be taken into account, we emphasize that due to the efficient iterative testing scheme of the PC algorithm, the dimension of 32 mentioned above refers to the maximum number of *parents* in the causal graph, not to the number of processes which can be much higher.

The first research question on causality, therefore, is easier to handle than consistently estimating a coupling strength using CMI. But also here we found that

the bias for low dimensional parents is not that severe. This demonstrates a great advantage of the two step approach because the assessment of coupling strength with MIT is only conducted using the small set of inferred parents and not the entire set of processes. Still, the quite novel estimators of CMI should be improved on and analytical distributions – or even just approximations thereof – would be a better alternative to expensive shuffle tests. The benefits of an MIT analysis for significance testing actually carry over to the general assessment of a well-interpretable coupling strength which is the subject of the next Chapter 5.

We also found that partial correlation can be used up to very high dimensions and with MIT the analytical significance test is reliable even for strongly autocorrelated time series which dramatically eases an analysis. This, also for CMI, provides an example of the statistical power gained using conditional inferences (Sect. 3.6). The gain in power essentially comes from the increase in degrees of freedom due to a proper conditioning. In Sect. 6.3, we will further discuss the differences between ITY and MIT on climatological examples. Table 4.1 also gives guidelines for the choice of parameters. Note that these conclusions were drawn only from the restricted setup studied here, especially for CMI it is hard to make general claims that hold for arbitrary distributions. Finally, note that nearest-neighbor estimators and also partial correlation require that the variables attain a continuous range of values which is the case for the climatological variables studied here, but other variables of interest, such as event time series, require different estimators. Continuous variables allow for a metric approach which is exploited by nearest-neighbor estimators, but not by conventional binning estimators, and can be seen as one reason why the former outperform the latter.

Table 4.1.: Conclusions from this chapter on limits of "well-behaved" estimation of CMI and partial correlation. D_Z is the dimension of conditions, T the time series length, and k the CMI estimation nearest-neighbor parameter. For tropical climate time series as studied in Chapter 6, the autocorrelation parameter a occurring in this table often is around or above 0.9. Note that the conclusions drawn here only hold for the set of models tested.

Property	CMI	Partial correlation
Detecting causality		
Good power as conditional independence test (AUC ≥ 0.8) studied for model Eq. (4.3)	$D_Z \leq 32$, $T \geq 1000$, $k = (0.05..0.5)\,T$	indep. of D_Z up to 64, $T \geq 100$
Impact of autocorrelation on detecting causality		
Actual false positives for $\alpha = 95\%$ due to autocorrelation $a = 0.9$ studied for model Eq. (4.7)	for $T = 1000$, $k = 50$: MI 80% ITY 35% MIT 15%	for $T = 100$: MI 30% ITY 10% MIT 5% (no error)
Maximum autocorrelation to limit false positive rate at $\approx 5\%$ studied for model Eq. (4.7)	MI $a < 0.6$ ITY $a < 0.8$ MIT $a < 0.9$	MI $a < 0.4$ ITY $a < 0.9$ MIT $a < 1$ (no error)
Quantifying causal strength		
Low bias and variance (relative MSE $\leq 30\%$) studied for model Eq. (4.3)	$D_Z \leq 8$, $T \geq 1000$, $k = 5..10$ indep. of T	indep. of D_Z up to 64, $T \geq 100$

Chapter 5.

Examples, theorems, and physical interpretation

5.1. Introduction – understanding measures

In Chapter 3, we have introduced measures to quantify causal interactions based on the notion of source entropy, but we have not yet enough justified why these measures are actually useful to physically understand interactions. The goal of this chapter is to develop a statistical and physical understanding of the different measures.

Since linear Gaussian processes are probably a good model for many aspects of climate interactions (Von Storch and Zwiers, 2002) and analytically tractable, in Sect. 5.2 we will compare the measures extensively on this Gaussian playground just like the harmonic oscillator is the toy model of quantum physics. We, again, discuss the problems of detecting causal relations and measuring a well-interpretable coupling strength separately. The former is extensively studied here regarding the impact of autocorrelation (Sect. 5.2.1) which are very common in climate time series. While the results from the previous chapter concerned the detection of the existence of causal links, here the emphasis lies on the detection of the correct coupling delay. We find that strong autocorrelations shift the peak of unconditional lag functions such as the cross correlation or mutual information which can lead to possibly misleading conclusions about the physical coupling delay. The problem of measuring causal strength is discussed here in a comparison of measures mostly coming from the physics literature, mutual information and transfer entropy (Sect. 5.2.2). We find that these measures are quite ambiguously influenced by autocorrelations and other external drivers while momentary information transfer excludes these effects making it well-interpretable. Several examples on the interaction along paths and between multiple processes in Sect. 5.2.3 and 5.2.4 enhance an intuition for these measures that will help in interpreting analyses of real climate data in Chapter 6. In Sect. 5.3, these results are substantiated by theorems proving the simple dependency of measures based in the idea of momentary information on the coefficients of nonlinear discrete-time stochastic processes. Numerical experiments in Sect. 5.4 further verify these results and extend the discussion on estimation from Chapter 4. In Sect. 5.5 we relate and interpret the interaction measures in the context of communication theory, thermodynamics, (geo-) physics, and network theory. The chapter is concluded by

a discussion of assumptions and limitations of measuring causal coupling strength (Sect. 5.6).

Some parts of this chapter contain results published in Runge et al. (2012b); Runge et al. (2014), while the analyses of interaction measures between multiple processes and much of the physical interpretations are new material.

5.2. Analytical examples

5.2.1. Pitfalls in inferring the delay and strength of a mechanism with cross correlations and regressions

In Sect. 2.2.2, we have reviewed the climate literature and gave a real data example of how coupling delays are frequently assessed. Also in many other fields of science interaction delays are inferred using the absolute maximum of the cross correlation or mutual information lag function (Sect. 3.4.1), which was proposed even in Granger and Lin (1994). Additionally, the correlation value at the maximum or regressions are used to evaluate the strength of an interaction. In the introduction in Sect. 2.2.2, we have asked how justified and reliable such an approach is and how the value of the cross correlation can be interpreted. Here, following Runge et al. (2014), we analyze a simple example to investigate how the value and lag at the maximum of the cross correlation function and regression coefficients depend on serial correlation or autocorrelation which is an ubiquitous feature of climate time series especially in the tropics (Von Storch and Zwiers, 2002) as also our motivational example in Sect. 2.2.2 has shown.

Consider the following bivariate first-order autoregressive process of two serially correlated subprocesses with a uni-directional influence of X on Y:

$$\begin{aligned} X_t &= aX_{t-1} + \eta_t^X \\ Y_t &= bY_{t-1} + cX_{t-1} + \eta_t^Y, \end{aligned} \tag{5.1}$$

where (η^X, η^Y) are independent and identically distributed Gaussian random variables, sometimes referred to as the *innovations*, with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}. \tag{5.2}$$

$|a|, |b| < 1$ is required for the process to be stationary while c can attain arbitrary (finite) values. In the following a, b – commonly regarded as the persistence of a process – are assumed positive as is often the case for climate time series (Von Storch and Zwiers, 2002).

Table 5.1.: Analytical comparison of lagged cross correlation and the partial correlation measures ITY and MIT as well as univariate and multivariate MIT regressions for the model example Eq. (5.1) on the dependent variable Y. The derivations can be found in Appendix A.2. The parents used in ITY and MIT for this model are $\mathcal{P}_{Y_t} = \{Y_{t-1}, X_{t-1}\}$ and $\mathcal{P}_{X_t} = \{X_{t-1}\}$. For the contemporaneous link $X_t - Y_t$, the MIT regression coefficient $B_{(X_{\mathcal{P}_X})_t}$ corresponds to the quotient of the residual's covariance and variance after lagged MIT regressions of X on its parents yielding $B_{X_{t-1}} = a$, and Y on its parents yielding $(B_{X_{t-1}}, B_{Y_{t-1}}) = (c, b)$. The formulas demonstrate the dependence of cross correlation and univariate regressions on the autocorrelations strengths a and b. Interestingly, for this model also ITY still depends on the autocorrelation strength of Y, while MIT fully excludes both autocorrelation influences.

Model example Eq. (5.1) for $\sigma_{XY} = 0$		Eq. (5.1) for $c = 0$
cross correlation $\rho(X_{t-\tau}; Y_t)$ $= \begin{cases} \frac{a^{1+\lvert\tau\rvert} c\sigma_X^2}{(1-a^2)(1-ab)} / \sqrt{\Gamma_X \Gamma_Y} & \text{for } \tau \leq 0 \\ \frac{c\sigma_X^2 \left(a^{\lvert\tau\rvert}(1-ab) - b^{\lvert\tau\rvert}(1-a^2)\right)}{(1-a^2)(a-b)(1-ab)} / \sqrt{\Gamma_X \Gamma_Y} & \text{for } \tau > 0 \end{cases}$ with $\Gamma_X = \frac{\sigma_X^2}{1-a^2}$, $\Gamma_Y = \frac{c^2\sigma_X^2(1+ab) + \sigma_Y^2(1-a^2)(1-ab)}{(1-a^2)(1-b^2)(1-ab)}$	univariate regression $B_{X_{t-1}} = \frac{c}{1-ab}$	univariate regression $B_{X_t} = \frac{(1-a^2)\sigma_{XY}}{(1-ab)\sigma_X^2}$
partial correlation $\rho^{\text{ITY}}_{X \to Y}(\tau)$ $= \begin{cases} \sqrt{\frac{c^2\sigma_X^2\left(c^2\sigma_X^2 + (1-ab)^2\sigma_Y^2\right)}{c^4\sigma_X^4 + 2(1-ab)c^2\sigma_X^2\sigma_Y^2 + (1-a^2)(1-ab)^2\sigma_Y^4}} & \text{for } \tau = 1 \\ 0 & \text{for } \tau \neq 1 \end{cases}$	MIT regression $\mathbf{B}_{\mathcal{P}_{\mathbf{Y_t}}} = \begin{pmatrix} B_{X_{t-1}} \\ B_{Y_{t-1}} \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix}$	MIT regression $B_{(X_{\mathcal{P}_X})_t} = \frac{\sigma_{XY}}{\sigma_X^2}$
partial correlation $\rho^{\text{MIT}}_{X \to Y}(\tau)$ $= \begin{cases} \frac{c\sigma_X}{\sqrt{c^2\sigma_X^2 + \sigma_Y^2}} & \text{for } \tau = 1 \\ 0 & \text{for } \tau \neq 1 \end{cases}$		

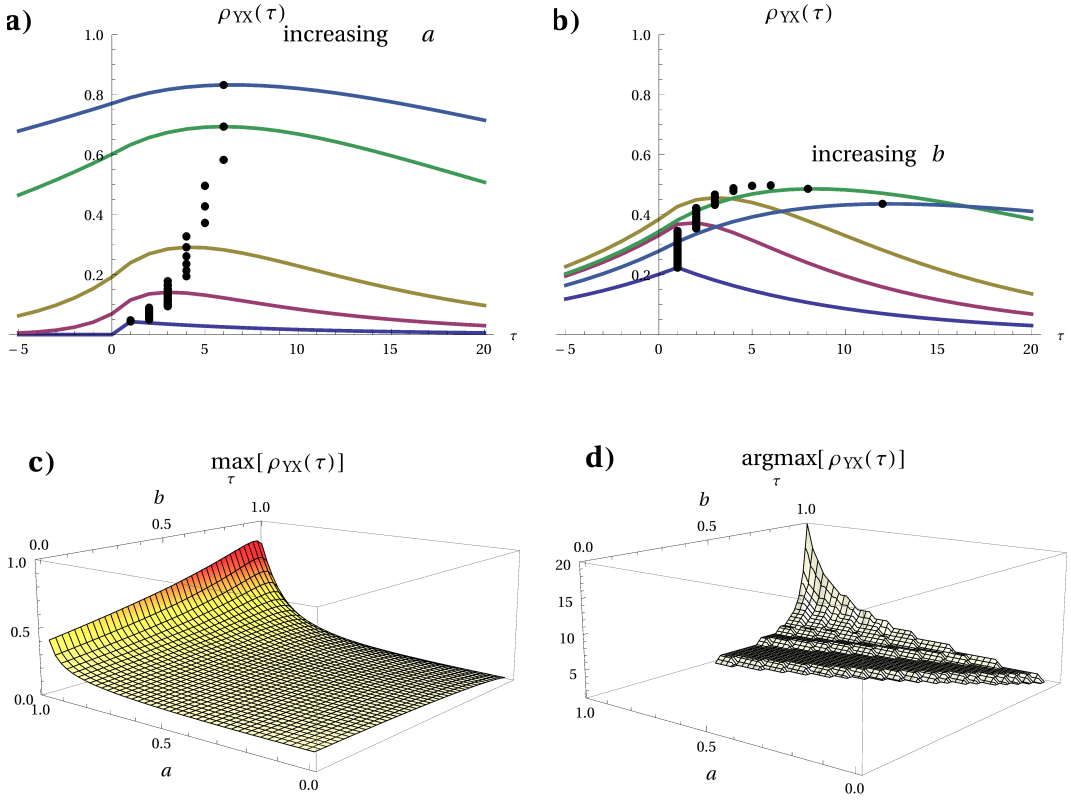


Figure 5.1.: Plots of the analytical cross correlation function given in Tab. 5.1 for model Eq. (5.1). (a) Correlation function for fixed $c = 0.1$, $\sigma_X = \sigma_Y = 1$, zero contemporaneous dependence $\sigma_{XY} = 0$, $b = 0.9$ and varying $a = 0, 0.6, 0.8, 0.95, 0.975$ (from bottom to top). The black dots indicate the maxima for the whole range from $a = 0$ to $a = 0.975$ in steps of 0.025. (b) Reverse case where $a = 0.9$ is fixed and b varies in the same range. (c) Value of the maximum and (d) the maximum's lag for varying a and b. In (d) only the region where the lag is shifted is plotted. In the model – assuming zero contemporaneous dependence $\sigma_{XY} = 0$ – this is independent of c the case if $b > \frac{1}{2}$ and $a > \frac{1-b}{b}$ assuming positive a and b. Note that the a-axis has been reversed for better visibility. The plots demonstrate the strong and nonlinear dependence of the correlation function on the autoregressive coefficients.

We now compare cross correlation and the partial correlations ITY and MIT and discuss the differences between univariate and MIT regressions for the model example Eq. (5.1).

Cross correlation lag function The cumbersome formula of the lagged cross correlation can be found in Tab. 5.1 in a comparison with the novel introduced partial correlation measures. Already from this formula we see that the correlation function $\rho_{YX}(\tau)$ clearly not only depends on c and the variances, but also on the autocorrelation coefficients a and b. To illustrate this dependence, we show in Figs. 5.1(a) and

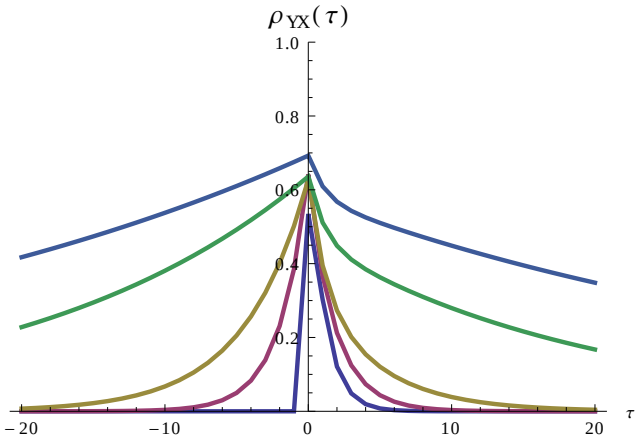


Figure 5.2.: Plot of the analytical cross correlation function for model Eq. (5.1) for fixed $c = 0.1$, $\sigma_X = \sigma_Y = 1$, contemporaneous dependence $\sigma_{XY} = 0.6$, $b = 0.4$ and varying $a = 0, 0.6, 0.8, 0.95, 0.975$ (from bottom to top). The plot shows that autocorrelation could even lead to a misinterpretation of the direction of influence.

(b) plots of $\rho_{YX}(\tau)$ for fixed coupling coefficient $c = 0.1$ and different autoregressive coefficients a while keeping $b = 0.8$ fixed in Fig. 5.1(a) and vice versa in Fig. 5.1(b).

Several observations can be made. The height of the peak of the correlation function for the same small coupling coefficient $c = 0.1$ strongly varies from very low to very high values for increasing autocorrelation strength a (Fig. 5.1(a). Especially for large autocorrelation, even a slight variation in a of 0.025 causes an increase in ρ of about 0.1. On the other hand, for increasing b the maximum first increases and for very large b decreases again (Fig. 5.1(b)) with an overall variation in ρ of about 0.3. Also the lag at which the maximum occurs is shifted towards larger lags for increasing a and b. This can happen even for low autocorrelations like $b = 0.6$, $a = 0.7$, while for tropical temperature anomalies values above 0.9 are very common as is the case in our motivating example (Sect. 2.2.2). Also here, for high autocorrelations, even for a slight variation in b of 0.025 the maximum's lag is shifted by up to 4. In Figs. 5.1(c) and (d) the value and lag of the maximum are plotted for all combinations of a and b. The maximum's value and lag are rather asymmetric and strongly nonlinear in their dependence on the coefficients. For increasing a the maximum can easily become very large and for additional large b the lag can be strongly shifted.

In Fig. 5.2, a case with additional contemporaneous covariance $\sigma_{XY} = 0.6$ is shown. Especially for the two upper curves, albeit the maximum is at lag zero, one is still tempted to interpret the larger correlation for negative lags as a sign for a mechanism where Y drives X, while actually the opposite is the case. Note that often interactions appear contemporaneous due to a low time resolution of the data, which can cause misleading physical interpretations.

As opposed to these complicated dependencies, those of the partial correlation

measures are much simpler. Firstly, ITY and MIT are non-zero only at the causal time lag $\tau = 1$. Regarding the value, we find that – counterintuitively – ITY actually still depends on the autocorrelation strength parameter b for this model, even though the past lag of Y is used as a condition. Only MIT fully excludes both influences and solely depends on the coupling coefficient c and the innovations' variances σ_X^2, σ_Y^2.

Regressions It is a common approach in regression analysis to regress Y on X at the lag with maximum correlation. As studied in the previous paragraph this can yield very misleading lags. Here two cases are studied: (i) Where the directional coupling coefficient is set $c = 0$, but the contemporaneous dependence σ_{XY} is non-zero. Then the maximum is at lag zero. (ii) with the contemporaneous covariance coefficient set to $\sigma_{XY} = 0$. Then, for moderately strong autocorrelation coefficients in model Eq. (5.1), i.e., outside the region shown in Fig. 5.1(d), the maximum will be at lag 1. For both cases the regression of Y_t on X_{t-1} and X_t, respectively, the coefficients $B_{X_{t-1}}$ and B_{X_t} can easily be derived from the covariances (given in Appendix A.2) demonstrating their dependence on a and b. The formulas are again shown in Tab. 5.1 for comparison. On the contrary, a multivariate MIT regression on the parents recovers the coefficients of the model, as shown in Tab. 5.1, without intermixing the coefficients as for the univariate regressions. This property can be proven to hold generally for autoregressive processes (Appendix A.6.2). In Fig. 5.3, the quotients of these coefficients divided by the coefficients $B'_{X_{t-1}}$ and B'_{X_t} for zero a and b are plotted for varying a and b to illustrate the factor by which the regression coefficient is changed due to autocorrelation. The plots show that the regression coefficient for lagged regressors varies nonlinearly in a and b and can be larger by orders of magnitude due to autocorrelation (Fig. 5.3(b)), while the regression coefficient for contemporaneous regressors can become zero (in the limit $a \to 1$) or even twice as

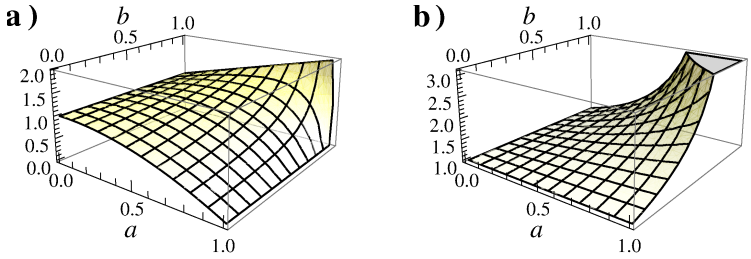


Figure 5.3.: Plots of quotients $B_{X_t}/B'_{X_t} = \frac{(1-a^2)}{(1-ab)}$ for a contemporaneous regression (case (i)) in (a) and $B_{X_{t-1}}/B'_{X_{t-1}} = \frac{1}{1-ab}$ for a lagged regression (case (ii)) in (b). These quotients describe the factor by which the regression coefficients are altered due to autocorrelations a and b. Note that the quotient $B_{X_{t-1}}/B'_{X_{t-1}}$ goes to infinity for $a = b = 1$. The plots demonstrate the strong nonlinear dependence of univariate regressions on autocorrelation.

large depending on a and b (Fig. 5.3(a)). Interestingly, for zero autocorrelation $a = 0$ in X, the autocorrelation b in Y makes no difference.

Summarizing, both the maximum's value and lag of the lagged cross correlation as well as regression coefficients are strongly affected by large autocorrelations and cannot be easily related to the coefficients of the underlying model. For high autocorrelations, these commonly applied measures are, therefore, not even a good first order approximation of the lag and coupling coefficient of the underlying model. The same qualitative behavior also holds for the mutual information lag function (for Gaussians, MI simply is a monotone transformation given by Eq. (3.36) and, thus, peaks at the same lag). This demonstrates the pitfalls of a typical coupling delay analysis as proposed by Granger and Lin (1994) and used in many fields as reviewed in Sect. 2.2.2. The causality detection algorithm introduced in Chapter 2 overcomes these problems because it unveils indirect links via the autodependencies within both processes that cause the shifting and increasing maximum.

Also regarding the maximum's value and regression coefficients, MIT and a multivariate regressions on the parents have a very simple dependency on the underlying model coefficients that makes them better interpretable than unconditional measures or univariate regressions. In the next section, we extend the comparison also to other measures defined in Section 3.4.

5.2.2. Comparison of measures of link strength

In this section, following Runge et al. (2012b), we compare mutual information (MI), transfer entropy (TE), the CMI defining causal links (LINK), information transfer to Y (ITY) and from X (ITX), and momentary information transfer (MIT) on an analytically tractable model of a multivariate Gaussian process:

$$\begin{aligned} Z_t &= c_{XZ} X_{t-1} + \eta_t^Z \\ X_t &= a_X X_{t-1} + \eta_t^X \\ Y_t &= c_{XY} X_{t-2} + c_{WY} W_{t-1} + \eta_t^Y \\ W_t &= \eta_t^W \end{aligned} \tag{5.3}$$

with independent Gaussian white noise processes $\eta_t^{\cdot}$ with variances $\sigma_{\cdot}^2$. The corresponding time series graph is depicted in Fig. 3.4 and the parents are $\mathcal{P}_{Y_t} = \{X_{t-2}, W_{t-1}\}$ and $\mathcal{P}_{X_{t-2}} = \{X_{t-3}\}$. Generally, the conditional entropy $H(Y|\mathbf{Z})$ of a D_Y-dimensional Gaussian process Y conditional on a (possibly multivariate) process $\mathbf{Z}$ is given by

$$H(Y|\mathbf{Z}) = \frac{1}{2} \ln \left((2\pi e)^{D_Y} \frac{|\Gamma_{Y\mathbf{Z}}|}{|\Gamma_{\mathbf{Z}}|} \right) \tag{5.4}$$

where $|\Gamma_{Y\mathbf{Z}}|$ is the determinant of the covariance matrix of $(Y, \mathbf{Z})$. In our case Y is univariate and thus $D_Y = 1$. The variances and covariances needed to evaluate

the determinants and detailed derivations for the following formulas are given in Appendix A.3.

First, we analyze TE given by Eq. (3.38). The TE between two components X, Y of a multivariate process $\mathbf{X} = (X, Y, Z, W)$ can be written as the difference of conditional entropies

$$I^{\mathrm{TE}}_{X\to Y} = H(Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-) - H(Y_t \,|\, \mathbf{X}_t^-), \tag{5.5}$$

where the latter entropy, conditioned on the whole infinite past, is actually the source entropy of Y and can be much easier computed by exploiting the Markov property

$$H(Y_t \,|\, \mathbf{X}_t^-) = H(Y_t|\mathcal{P}_{Y_t}), \tag{5.6}$$

which yields, using Eq. (5.4),

$$\begin{aligned} H(Y_t|\mathcal{P}_{Y_t}) &= \frac{1}{2}\ln\left(2\pi e \frac{|\Gamma_{Y_t X_{t-2} W_{t-1}}|}{|\Gamma_{X_{t-2},W_{t-1}}|}\right) \\ &= \frac{1}{2}\ln\left(2\pi e \sigma_Y^2\right). \end{aligned} \tag{5.7}$$

The source entropy of Y is therefore given by the entropy of the innovation term η^Y. In the first entropy term, on the other hand, the infinite vector cannot be treated that easily and we have to evaluate the determinants of infinite dimensional matrices in

$$H(Y_t|Y_t^-, W_t^-, Z_t^-) = \frac{1}{2}\ln\left(2\pi e \frac{|\Gamma_{Y_t Y_t^- W_t^- Z_t^-}|}{|\Gamma_{Y_t^- W_t^- Z_t^-}|}\right). \tag{5.8}$$

However, for the special case of $c_{XZ} = c_{WY} = 0$, i.e., no input processes apart from the autodependency in X, the quotient of these matrix determinants can be simplified to the quotient of infinite Toeplitz matrix determinants. As shown in Appendix A.3.1, we can then apply Szegö's theorem (Szegö, 1915; Böttcher et al., 2006) and get

$$I^{\mathrm{TE}}_{X\to Y} \overset{c_{XZ}=c_{WY}=0}{=} \frac{1}{2}\ln\left(1 + \frac{(c_{XY}^2\sigma_X^2)/(1-a_X^2)}{\sigma_Y^2}\right). \tag{5.9}$$

Another tractable case is $a_X = 0$ for which the blocks of the covariance matrix $\Gamma_{Y_t Y_t^- W_t^- Z_t^-}$ become diagonal and

$$I^{\mathrm{TE}}_{X\to Y} \overset{a_X=0}{=} \frac{1}{2}\ln\left(1 + \frac{c_{XY}^2\sigma_X^2\sigma_Z^2}{\sigma_Y^2(c_{XZ}^2\sigma_X^2 + \sigma_Z^2)}\right). \tag{5.10}$$

Thus, in the first case the value of TE for our model depends on the autodependency coefficient and in the second case on the coupling coefficient and variance of Z. But why should a measure of coupling strength between X and Y depend on internal

dynamics of X and, even more so, on the interaction of X with another process Z? While it can be information-theoretically explained, it seems rather unintuitive for a measure of coupling strength between X and Y.

Next, we compute the CMI $I^{\text{LINK}}_{X\to Y}$ that defines links in a time series graph. Writing Eq. (3.43) for $\tau = 2$ as a difference of conditional entropies, the second term is again the source entropy as given by Eq. (5.7) and in this case also the first entropy can be simplified using the Markov property

$$H(Y_t \,|\, \mathbf{X}_t^- \backslash X_{t-2}) = H(Y_t \,|\, \mathbf{X}_t^{(t-1,\,\ldots,\,t-3)} \backslash \{X_{t-2}\}) \tag{5.11}$$

to arrive at a finite covariance matrix from which a lengthy computation yields

$$I^{\text{LINK}}_{X\to Y} = \frac{1}{2}\ln\left(1 + \frac{c_{XY}^2\sigma_X^2\sigma_Z^2}{\sigma_Y^2(c_{XZ}^2\sigma_X^2 + (1{+}a_X^2)\sigma_Z^2)}\right). \tag{5.12}$$

Again, also this measure of coupling strength depends on the coefficients belonging to other coupling and autodependency links.

We now turn to the measures that solely use the parents as conditions which has the analytical and numerical advantage of low dimensional computations. The resulting expressions for the CMI with no conditions, i.e., the mutual information (MI), and for either one of the parents as a condition for $\tau = 2$ are

$$I^{\text{MI}}_{X\to Y} = \frac{1}{2}\ln\left(1 + \frac{(c_{XY}^2\sigma_X^2)/(1{-}a_X^2)}{c_{WY}^2\sigma_W^2 + \sigma_Y^2}\right), \tag{5.13}$$

$$I^{\text{ITY}}_{X\to Y} = \frac{1}{2}\ln\left(1 + \frac{(c_{XY}^2\sigma_X^2)/(1{-}a_X^2)}{\sigma_Y^2}\right), \tag{5.14}$$

$$I^{\text{ITX}}_{X\to Y} = \frac{1}{2}\ln\left(1 + \frac{c_{XY}^2\sigma_X^2}{c_{WY}^2\sigma_W^2 + \sigma_Y^2}\right). \tag{5.15}$$

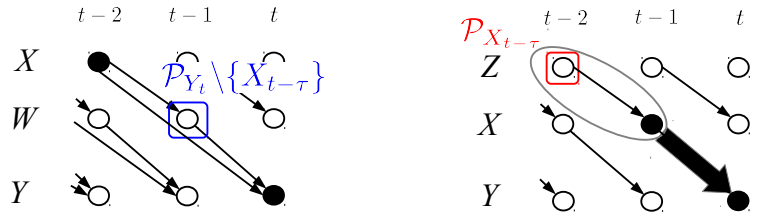
Thus MI depends on the coefficients and variances of the input processes, while ITX and ITY still depend at least on the coefficient and variance of the process that is not conditioned on. Contrary to TE and LINK though, neither of the three measures depends on the interaction with Z. In our model the inputs to X and Y, i.e., the autodependency with X_{t-3} and the external input from W_{t-1}, are independent which makes the formulas much simpler.

Finally, the MIT for $\tau = 2$ is

$$I^{\text{MIT}}_{X\to Y} = \frac{1}{2}\ln\left(1 + \frac{c_{XY}^2\sigma_X^2}{\sigma_Y^2}\right), \tag{5.16}$$

which solely depends on the model coefficients that govern the source entropies, i.e., the variances σ_X^2, σ_Y^2, and the coupling coefficient c_{XY}.

This equation can be proven to hold for arbitrary multivariate linear autoregressive processes (Sect. A.6.2). More generally, for a class of additive models MIT depends



a) Sidepath **b)** Nonlinear coupling

Figure 5.4.: Two examples of couplings that cannot be related to one single coefficient c_{XY}. Black dots mark $X_{t-\tau}$ and Y_t, the red and blue boxes their parents. (a) Sidepath, i.e., if there exists a path from X_{t-2} to some parent of Y_t. Then the coupling cannot be related to one single link, but additionally to the path via W_{t-1}. (b) Visualization of a nonlinear coupling discussed later in Sect. 5.2.5 between X_{t-1} and Y_t. In this case the entropies of X_{t-1} and its parents "mix" and the coupling could be considered as emanating from $(X_{t-1}, \mathcal{P}_{X_{t-1}})$ rather than X_{t-1} alone.

only on the coupling coefficient c_{XY} and the source variances of η^X and η^Y as will be proven in the coupling strength autonomy theorem in Section 5.3.2.

5.2.3. Interactions along paths

In the previous Sect. 5.2.2, we studied the interactions between two processes and quantified the strength of causal links. The main finding was that MIT solely depends on the coefficient corresponding to the causal link. But the following – still linear – example model visualized in Fig. 5.4(a) shows that a coupling mechanism cannot always be associated with one single coupling coefficient c_{XY}:

$$\begin{aligned} X_t &= \eta_t^X \\ W_t &= c_{XW} X_{t-1} + \eta_t^W \\ Y_t &= c_{XY} X_{t-2} + c_{WY} W_{t-1} + \eta_t^Y \end{aligned} \tag{5.17}$$

where the influence of X_{t-2} on Y_t has two paths: One via the direct coupling link "$X_{t-2} \to Y_t$" and one via the path "$X_{t-2} \to W_{t-1} \to Y_t$" such that we can rewrite

$$Y_t = c_{XY} X_{t-2} + c_{WY}(c_{XW} X_{t-2} + \eta_{t-1}^W) + \eta_t^Y , \tag{5.18}$$

from which we see, that the coupling cannot be unambiguously related to one coefficient. Here, MIT at $\tau = 2$ is

$$I^{\mathrm{MIT}}_{X \to Y} = \frac{1}{2} \ln \left(1 + \frac{c_{XY}^2 \sigma_X^2 \sigma_W^2}{\sigma_Y^2 (c_{XW}^2 \sigma_X^2 + \sigma_W^2)} \right) , \tag{5.19}$$

and depends not only on c_{XY}, but also on the coefficient c_{XW} of the link "$X_{t-2} \to W_{t-1}$", and on the variance of W. In this case it might be more appropriate to "leave open" both paths and exclude W_{t-1} from the conditions arriving at the momentary information transfer along paths (MITP) defined in Eq. (3.53) at $\tau = 2$

$$I^{\text{MITP}}_{X \to Y} = \frac{1}{2} \ln \left(1 + \frac{(c_{XY} + c_{XW} c_{WY})^2 \sigma_X^2}{c_{WY}^2 \sigma_W^2 + \sigma_Y^2} \right). \tag{5.20}$$

Here the sum $c_{XY} + c_{XW} c_{WY}$ is the covariance along both paths, which can also vanish for $c_{XY} = -c_{XW} c_{WY}$ (a pathological case where the causal assumption of faithfulness is violated as discussed, see also Sect. 2.4.7), and seems like a more appropriate representation of the coupling between X_{t-2} and Y_t. While in this example there are no external parents influencing the processes along the path, in more complex schemes as shown in Fig. 3.6 their effect can be excluded by the condition on the parents of the nodes on the path denoted by $\mathcal{C}_{X_{t-\tau} \to Y_t}$. In Sect. 5.3.2 this will be proven for the general case.

For the linear framework there exists a whole theory of quantifying the relative influence of paths between two processes (Wright, 1934). In *path analysis* the weight of each link is assessed by the standardized multiple regression coefficient (beta coefficient) assuming a certain connectivity. Then one can decompose the cross correlation between two processes as the sum over all open path weights, where the path weight is given by the product of the link weights. We can use the time series graph as the assumed connectivity model and estimate the link coefficients by a multiple MIT regression on the parents. Then the correlation decomposition allows to use the 'locally' estimated weights as a measure of the global correlation between two processes. This framework rests, however, on a specified model yielding coefficients or a structural equation model (Spirtes et al., 1998) in a more general setting. In the information-theoretic framework such a decomposition seems to be impossible (albeit we hope that there is a way). Therefore, in the next section we will explore how interaction information can give similar insights into the effect of an intermediate process on a causal path between two others.

In Sect. 6.5, we discuss how MITP can be used to quantify the influence of momentary perturbations entering the system at X on causally non-adjacent nodes in the time series graph in a climatological application.

5.2.4. Interactions between multiple processes

Here, we discuss the measures introduced in Section 3.5 that quantify the interaction between three nodes along paths. In Fig. 5.5(a) we show the simplest example motif of three processes interacting causally via directed links. In Appendix A.5, we discuss all four motifs that include all possible combinations of causal and contemporaneous links between three processes. Essentially, these represent the four unconditioned open motifs defined in Fig. 2.5. We consider the momentary interaction information

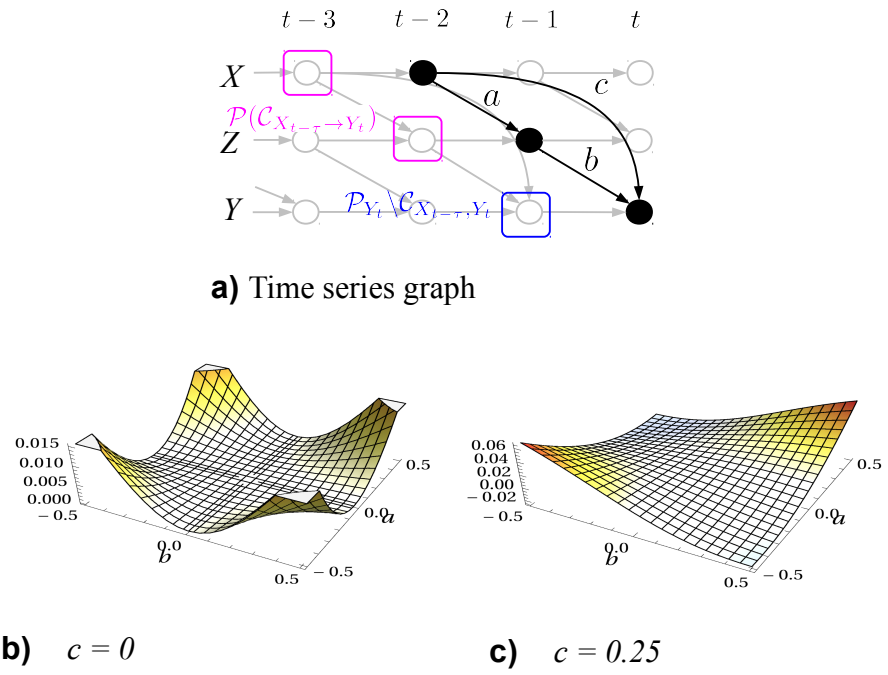


Figure 5.5.: (a) Time series graphs for the simplest example of a causal interaction between three processes (black dots). The parents of Y_t and the path nodes $\mathcal{C}_{X_{t-\tau}\to Y_t} = \{X_{t-\tau}, Z_{t-\tau_Z}\}$ are shown as blue and magenta boxes, respectively. a, b, c denote the coefficients for the Gaussian example analyzed in the text. (b,c) Momentary interaction information (MII) for the Gaussian example given by Eq. (5.22) for varying a, b and fixed $c = 0$ (b) and $c = 0.25$ (c) and unit variances. Note that in (c) MII can become negative (blue shading).

(MII, Eq. (3.55)) of the three black nodes in Fig. 5.5(a). For this special causal case of three processes we can evaluate MII using the relation

$$\mathcal{I}^{\text{MII}}_{X\to Y|Z} = I^{\text{MITP}}_{X\to Y} - I^{\text{MIT}}_{X\to Y}. \tag{5.21}$$

If all processes are jointly Gaussian with coefficients denoted in the figure and innovation's variances σ^2_X, σ^2_Y, σ^2_Z, their MII can be evaluated by inserting Eqns. (5.19) and (5.20) – renaming the coefficients – giving

$$\mathcal{I}^{\text{MII}}_{X\to Y|Z} = \frac{1}{2}\ln\left(\frac{(\sigma^2_Y + c^2\sigma^2_X + (ac+b)^2\sigma^2_Z)(\sigma^2_X + a^2\sigma^2_Z)}{((\sigma^2_Y + b^2\sigma^2_Z)(\sigma^2_X + a^2\sigma^2_Z) - a^2b^2\sigma^4_Z)} \times \frac{\sigma^2_Y}{(\sigma^2_Y + c^2\sigma^2_X)}\right), \tag{5.22}$$

as derived in Appendix A.5.

Just like in the examples of Sect. 5.2.2 for MIT, here MII only depends on the coefficients between the interacting processes and excludes information from the past due to autocorrelations or other dependencies. Also for MII this feature can be

proven for a general class of processes as shown in Sect. 5.3. It allows to isolate the interactions of interest from the rest of the complex system.

Formula (5.22) is plotted for $c = 0$ and varying a, b, and unit variances in Fig. 5.5(b). MII is always positive because the only interaction stems from the path and $I^{\text{MIT}}_{X\to Y}$ is zero demonstrating the explanatory influence of Z, which acts as a mediating process. In the Venn diagram of Fig. 3.2(c) this corresponds to the case that $H(Z)$ entails all of the shared entropy between X and Y.

The case $c = 0.25$ is plotted in Fig. 5.5(c). Then MII can become negative if a and b are of different sign. This constitutes the interesting case, that an anticorrelated coupling mechanism along the chain $X \to Z \to Y$ counteracts a positive direct coupling mechanism $X \to Y$ leading to a reduced net influence of X on Y. We will see in the applications, that this is a common mechanism in climate (Sect. 6.5.5). If, on the other hand, a and b are of the same sign, both mechanisms act in concordance and enhance the influence of X on Y. For negative c, the case is reversed, i.e., a and b of equal sign counteract and vice versa. If both a and b are zero, the three processes are not causally linked anymore and MII is zero.

As shown in Appendix A.5, a process Z can only affect the interaction between X and Y if it is an intermediate process on a causal directed path, i.e., if it is in the set $\mathcal{C}_{X_{t-\tau}\to Y_t}$. Due to the symmetry of interaction information the interactions can also be written as

$$\mathcal{I}^{\text{MII}}_{X\to Y|Z} = \mathcal{I}^{\text{MII}}_{Z\to Y|X}, \tag{5.23}$$

while the last case $\mathcal{I}^{\text{MII}}_{Z\to X|Y}$ does not pertain to a causal interaction. Then our findings imply that also a common driver X for the interaction between Z and Y can enhance or counteract. This, however affects only measures that do not exclude the parents of *both* processes like mutual information. The MITP between Z and Y, for example, is conditioned on X and excludes possible interactions.

Our examples indicate that also MII has very simple dependencies solely on the coefficients along the coupling mechanisms by which the three processes interact. For the case of causal triples without other paths, MII is the interaction of the source entropy of X with Z and Y. In Sect. 5.5.4 we discuss how MII can be used as a measure of 'causal interaction betweenness', complementing concepts from complex network theory. This discussion will help to interpret the interactions analyzed in climate applications in Sect. 6.5.

5.2.5. Nonlinear dependencies

Another example where one cannot unambiguously relate the coupling strength to one coefficient is for a nonlinear dependency between X and Y (Fig. 5.4(b)):

$$\begin{aligned} Z_t &= \eta^Z_t \\ X_t &= c_{ZX} Z_{t-1} + \eta^X_t \\ Y_t &= c_{XY} (X_{t-1})^2 + \eta^Y_t . \end{aligned} \tag{5.24}$$

If we express Y_t explicitly in terms of the source variance of X and the parent of X,

$$Y_t = c_{XY} c_{ZX}^2 Z_{t-2}^2 + 2 c_{ZX} c_{XY} Z_{t-2} \eta_{t-1}^X + c_{XY} (\eta_{t-1}^X)^2 + \eta_t^Y \,, \tag{5.25}$$

we note that due to the term $2 c_{ZX} c_{XY} Z_{t-2} \eta_{t-1}^X$ the effect of Z_{t-2} is not additively separable from the source process η_{t-1}^X. Figure 5.6 shows the scatter plot of a realization. One can see, that the functional form of the dependency between X and Y varies with Z and can, thus not be conditioned out. In the Venn diagram of Fig. 3.5(b) this "mixing" of entropies implies that the parts of the entropies $H(X|\mathcal{P}_X)$ and $H(\mathcal{P}_X)$ that overlap with $H(Y)$ are not distinguishable anymore, which could be visualized by the red and light gray shadings bleeding into one another. Therefore the coupling could be considered as emanating from $(X_{t-1}, \mathcal{P}_{X_{t-1}})$ rather than X_{t-1} alone (visualized by a thick arrow in Fig. 5.4(b)). For this nonlinear model we have not found an analytical expression for MIT, but the more general case of this model is studied numerically in the Sect. 5.4. Here a multivariate CMI like $I^{\mathrm{SITY}}_{X \to Y}$ (Eq. (3.60)) could be a more appropriate measure of coupling strength.

This example of nonlinear dependencies points to constraints under which full coupling strength autonomy can be reached. In Section 5.3, we will formalize these constraints to general conditions in a theorem of coupling strength autonomy.

5.2.6. Decomposing covariance as a superposition of paths

Before moving on to generalizations of the results found for MIT and MII in the last sections, we briefly show how mutual information (MI) in the linear case can be

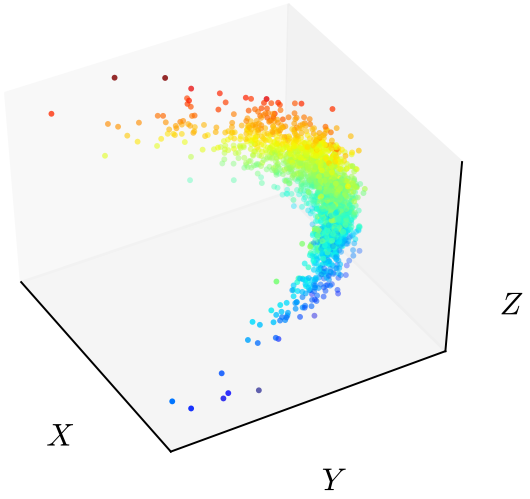


Figure 5.6.: Scatter plot of a realization of the nonlinear example model Eq. (5.24). The colors scale with the value of Z.

viewed as a superposition of entropy paths following the ideas in Wright (1934). For linear multivariate Gaussian processes, MI is given by (see Eq. 3.36)

$$I^{\mathrm{MI}}_{\mathrm{Gauss}}(X_{t-\tau};Y_t) = -\frac{1}{2}\ln\left(1 - \frac{E[X_{t+\tau}Y_t]^2}{|E[Y_tY_t]|\cdot|E[X_tX_t]|}\right), \tag{5.26}$$

where $E[\cdots]$ denotes the expectation value.

For an autoregressive process given by Eq. (2.14) there exists an analytical expression of the lagged covariance in terms of Φ (Brockwell and Davis, 2009, Ch. 11.3):

$$\Gamma_{ij}(\tau) \equiv E[\mathbf{X}^i_{t+\tau}\mathbf{X}^j_t] = \sum_{n=0}^{\infty}\Big(\Psi(n+\tau)\Sigma\Psi^\top(n)\Big)_{ij} \tag{5.27}$$

where $\Psi(n)$ can be recursively computed from matrix products:

$$\Psi(n) \equiv \sum_{s=1}^{n}\Phi(s)\Psi(n-s), \tag{5.28}$$

for example,

$$\begin{aligned}
&\Psi(0) = \mathbb{I},\\
&\Psi(1) = \Phi(1),\\
&\Psi(2) = \Phi^2(1) + \Phi(2),\\
&\Psi(3) = \Phi^3(1) + \Phi(1)\Phi(2) + \Phi(2)\Phi(1) + \Phi(3),
\end{aligned} \tag{5.29}$$

where $\mathbb{I}$ is the identity matrix.

Now, as a non-zero entry in Φ corresponds to a link, graph-theoretically an entry $\Psi(3)_{ij} \neq 0$ can be interpreted as a superposition of the contributions from different paths in the time series graph, each with total delay 3: one direct path of only one link with lag 3 ($\Phi(3)_{ij}$), paths composed of two links where the first has lag 1 and the second lag 2 ($(\Phi(1)\Phi(2))_{ij}$) and vice versa ($(\Phi(2)\Phi(1))_{ij}$), and paths comprised of three links, each with lag 1 ($(\Phi^3(1))_{ij}$). The covariance $\Gamma_{YX}(\tau)$ and, consequently the mutual information and cross correlation, then is an infinite sum of the triple product of matrix powers comprised of the coefficient and innovation's covariance matrix and therefore a nonlinear polynomial combination of coefficients of *all possible paths* that end in X and τ-lags later in Y, emanating from nodes and their contemporaneous neighbors at all possible lags. These paths can be read off from the time series graph. In essence, most spurious links in the mutual information lag function are due to the common driver effect of past lags (Fig. 2.3(b)) or the indirect causal effect due to intermediate lags (Fig. 2.3(a)). In Appendix A.6, we give another decomposition of covariance in terms of the parents and discuss further results for the linear theory.

5.3. Theorems

5.3.1. Causality theorem (Markov property)

By the Markov property Eq. (2.12), all measures between two processes X and Y listed in Tab. 3.1 that include the parents of Y in their conditions (i.e., LINK, ITY, MIT) can be called causal measures in that they are zero if X is not a parent of Y. This holds for arbitrary functional dependencies between Y and its parents as proven in the following theorem. Similar theorems hold for TE, DTE, and SITY as discussed below.

Theorem 5.1 (Causality)**.** *For a general discrete-time stochastic process $\mathbf{X}$, let $Y \in \mathbf{X}$ be a univariate subprocess depending on its parents $\mathcal{P}_{Y_t} \subset \mathbf{X}_t^-$ by*

$$Y_t = f(\mathcal{P}_{Y_t}, \eta_t^Y), \tag{5.30}$$

where f is an arbitrary function and η_t^Y is a stochastic process with some arbitrary distribution, that is independent of the past of $\mathbf{X}$, i.e., $\eta_t^Y \perp\!\!\!\perp \mathbf{X}_t^-$, and independent in time, i.e., $\eta_t^Y \perp\!\!\!\perp \eta_{t'}^Y$ for $t' \neq t$. Now we consider the CMI $I(X_{t-\tau}; Y_t|\mathcal{S})$ with $X \in \mathbf{X}$, $\tau > 0$, and a set of conditions $\mathcal{S}$ that entails the parents, i.e., $\mathcal{P}_Y \subseteq \mathcal{S} \subset \mathbf{X}_t^-$. Then

$$X_{t-\tau} \notin \mathcal{P}_{Y_t} \implies I(X_{t-\tau}; Y_t|\mathcal{S}) = 0. \tag{5.31}$$

Proof. Using the data processing inequality of CMI (Eq. (3.23)), translational invariance (Eq. (3.26) and the fact that the joint entropy with a constant reduces to the marginal entropy:

$$\begin{aligned}
&I(X_{t-\tau}; Y_t|\mathcal{S}) \\
&= I(X_{t-\tau}; f(\mathcal{P}_{Y_t}, \eta_t^Y)|\mathcal{S}) \\
&\leq I(X_{t-\tau}; (\mathcal{P}_{Y_t}, \eta_t^Y)|\mathcal{S}) \quad \text{(data processing inequality (Eq. (3.23)))} && (5.32)\\
&= I(X_{t-\tau}; \mathcal{P}_{Y_t}|\mathcal{S}) + \underbrace{I(X_{t-\tau}; \eta_t^Y|\mathcal{S}, \mathcal{P}_{Y_t})}_{=0 \text{ (i.i.d. noise)}} \quad \text{(chain rule for CMI (Eq. (3.22)))} && (5.33)\\
&= I(X_{t-\tau}; \mathcal{P}_{Y_t}|\mathcal{S}) = 0 \quad \text{(translational invariance (Eq. (3.26)))} && (5.34)
\end{aligned}$$

□

For $\mathcal{S} = \mathbf{X}_t^-$, this proves causality for LINK, with $\mathcal{S} = \mathcal{P}_{Y_t}$ for ITY, and with $\mathcal{S} = (\mathcal{P}_{Y_t}, \mathcal{P}_{X_{t-\tau}})$ for MIT. A similar theorem holds for SITY, where we have to demand that $(X_{t-\tau}, \mathcal{P}^X_{X_{t-\tau}})$ is not part of the parents. For TE and DTE the same holds if X_t^- is not part of $\mathbf{X}_t^-$.

The other direction, $I(X_{t-\tau}; Y_t|\mathcal{S}) = 0 \implies X_{t-\tau} \notin \mathcal{P}_{Y_t}$ does not hold if, for example, the influence of $X_{t-\tau}$ is counteracted by a side path resulting in a vanishing shared information (a pathological case where the causal assumption of faithfulness is violated, see also Sect. 2.4.7).

For the case of the path-based measures ITX, ITP and MITP the property 'causal' in Tab. 3.1 is defined in the wider sense that $X_{t-\tau}$ and Y_t are not necessarily directly linked, but connected by a directed causal path as defined in Section. 2.4.3. Consequently also the interaction measures IIP and MII are causal because they are derived from these path-based measures.

To prove theorems regarding the second research question on causal strength, we, unfortunately, need to make more assumptions on the functional dependence of Y on its parents.

5.3.2. Coupling strength autonomy

Momentary information transfer (MIT)

In this section, following Runge et al. (2012b), we generalize the examples discussed in Sections 5.2.2, 5.2.3 and 5.2.4.

Let X, Y be two subprocesses of some multivariate stationary discrete-time process $\mathbf{X}$ sufficing the Markov property Eq. (2.12) (Spirtes et al., 2000; Pearl, 2000) with time series graph $\mathcal{G}$ and coupling link "$X_{t-\tau} \to Y_t$" for $\tau > 0$. The following derivations also hold for more than one link at lags $\tau' \neq \tau$ between X and Y. As before, we denote their parents $\mathcal{P}_{Y_t}$ and $\mathcal{P}_{X_{t-\tau}}$. For the link "$X_{t-\tau} \to Y_t$" we define the following conditions:

1. *Additivity* means that the dependence of X_t on its source process η_t^X and parents $\mathcal{P}_{X_t}$ and of Y_t on its source process η_t^Y, $X_{t-\tau}$ and the remaining parents $\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\}$ is *additive*, i.e., they can be written as

   $$\begin{aligned} X_t &= g_X(\mathcal{P}_{X_t}) + \eta_t^X \\ Y_t &= f(X_{t-\tau}) + g_Y(\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\}) + \eta_t^Y \end{aligned} \tag{5.35}$$

   for possibly multivariate random variables $\mathcal{P}_{X_t}$ and $\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\}$, univariate i.i.d. random variables η^X and η^Y with arbitrary, not necessarily identical distributions, and arbitrary functions g_Y, g_X, f.

2. *Linearity in f:* The dependence of Y_t on $X_{t-\tau}$ is linear, i.e.,

   $$f(x) = cx \tag{5.36}$$

   with real c.

3. *"No sidepath"-condition*: In the time series graph $\mathcal{G}$ the node $X_{t-\tau}$ is separated from $(\mathcal{P}_{Y_t} \setminus \mathcal{P}_{X_{t-\tau}}) \setminus \{X_{t-\tau}\}$ given $\mathcal{P}_{X_{t-\tau}}$ (for a formal definition of paths and separation see Sect. 2.4.3). Due to the Markov property (Eq. (2.12)) this separation implies that

   $$I((\mathcal{P}_{Y_t} \setminus \mathcal{P}_{X_{t-\tau}}) \setminus \{X_{t-\tau}\}; X_{t-\tau} | \mathcal{P}_{X_{t-\tau}}) = 0. \tag{5.37}$$

The latter constraint essentially assumes that there are no links or directed paths from X or its contemporaneous neighbors to any other parents of Y. This would imply, that in the Venn diagram of Fig. 3.5(b) the light grey and blue parts overlap.

Theorem 5.2 (Inequality relations between ITX, MIT, and ITY)**.** *For $\tau > 0$ and under the "no sidepath"-condition Eq. (5.37), ITX, MIT and ITY are related by the inequality*

$$I^{\mathrm{ITX}}_{X\to Y}(\tau) \leq I^{\mathrm{MIT}}_{X\to Y}(\tau) \leq I^{\mathrm{ITY}}_{X\to Y}(\tau). \tag{5.38}$$

This holds independent of linearity and even additivity. The right part of the inequality also holds for sidepaths. The proof is given in Appendix A.4.1. From this inequality it also follows that the interaction information $\mathcal{I}(X_{t-\tau}; Y_t; \mathcal{P}_{X_{t-\tau}}|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\})$ is always positive. It quantifies how much the parents of $X_{t-\tau}$ enhance the interaction with Y_t as discussed in Sect. 3.5.2.

Theorem 5.3 (Coupling strength autonomy for MIT)**.** *MIT defined in Eq. (3.48) for the coupling link "$X_{t-\tau} \to Y_t$" for $\tau > 0$ of a multivariate stationary discrete-time process $\mathbf{X}$ sufficing the Markov property has the following dependency properties:*

1. *If all three conditions (5.35, 5.36, 5.37) hold, then MIT can be expressed as an MI of the source processes:*

   $$I^{\mathrm{MIT}}_{X\to Y}(\tau) = I(\eta^X_{t-\tau}; c\eta^X_{t-\tau} + \eta^Y_t). \tag{5.39}$$

   Since η^Y_t and $\eta^X_{t-\tau}$ are assumed to be independent, the probability density of their sum is given by their convolution. The MIT thus depends solely on c and the joint and marginal distributions of $\eta^X_{t-\tau}$ and the convolution of η^Y_t with $c\eta^X_{t-\tau}$.

2. *If only conditions (5.35, 5.36) hold, i.e., there exists a sidepath between $X_{t-\tau}$ and some nodes in $\mathcal{P}_{Y_t} \setminus \mathcal{P}_{X_{t-\tau}}$, then MIT depends additionally on the distributions of at least the "sidepath-parents" in $\mathcal{P}_{Y_t}$ and their functional dependence on Y_t:*

   $$I^{\mathrm{MIT}}_{X\to Y}(\tau) = I(\eta^X_{t-\tau}; c\eta^X_{t-\tau} + \eta^Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}). \tag{5.40}$$

   This relation can be further simplified if $g_Y(\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\})$ is additive in some parents.

3. *If only the additivity condition (5.35) holds, i.e., $f(x)$ is nonlinear and mixes $\eta^X_{t-\tau}$ with the parents $\mathcal{P}_{X_{t-\tau}}$ then MIT depends additionally on f, the distributions of variables in $\mathcal{P}_{X_{t-\tau}}$ as well as $\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}$ and their functional dependencies on Y_t:*

   $$I^{\mathrm{MIT}}_{X\to Y}(\tau) = I(\eta^X_{t-\tau};\, f(\eta^X_{t-\tau} + g_X(\mathcal{P}_{X_{t-\tau}})) + \eta^Y_t \mid \mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\},\, \mathcal{P}_{X_{t-\tau}}). \tag{5.41}$$

   This relation can be further simplified if some parents in $\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}$ are independent of $f(\eta^X_{t-\tau} + g_X(\mathcal{P}_{X_{t-\tau}}))$.

For a contemporaneous link "$X_t - Y_t$" the contemporaneous MIT defined in Eq. (3.49) under the condition (5.35) is:

$$I^{\text{MIT}}_{X-Y} = I(\eta^X_t; \eta^Y_t | \mathcal{N}_{X_t} \backslash \{Y_t\}, \mathcal{N}_{Y_t} \backslash \{X_t\}). \tag{5.42}$$

A contemporaneous link cannot have sidepaths. For $X = Y$, MIT measures the autodependency strength.

The proofs are given in Appendix A.4.2. We discuss the implications of this theorem at the end of this section.

Linear case For the special case of multivariate linear autoregressive processes of order p (Brockwell and Davis, 2009) defined by Eq. (2.14),

$$\mathbf{X}_t = \sum_{s=1}^{p} \mathbf{\Phi}(s)\mathbf{X}_{t-s} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \tag{5.43}$$

with the coupling coefficient c_{XY} at lag τ corresponding to the connectivity matrix entry $\mathbf{\Phi}(\tau)_{YX}$, and with no sidepaths, Eq. (5.39) leads to

$$I^{\text{MIT}}_{X \to Y}(\tau) = \frac{1}{2} \ln \left(1 + \frac{c^2_{XY} \sigma^2_X}{\sigma^2_Y} \right), \tag{5.44}$$

generalizing the MIT for our analytical model in Eq. (5.16). For an autodependency at lag τ with coefficient a_Y and no sidepaths the MIT is $I^{\text{MIT}}_{Y \to Y}(\tau) = \frac{1}{2} \ln (1 + a^2_Y)$, independent of the source variance σ^2_Y. The linear contemporaneous MIT is equal to the partial correlation of the residuals as follows from Eq. (5.42).

Momentary information transfer along paths (MITP)

For the path-based MITP, we can also prove simple dependencies. We assume a linear dependence of Y on the path nodes $\mathcal{C}_{X_{t-\tau} \to Y_t}$ including $X_{t-\tau}$ and for all other dependencies only additivity.

Theorem 5.4 (Coupling strength autonomy for MITP)**.** *Let X, Y be two subcomponents of a multivariate stationary discrete-time process $\mathbf{X}$ sufficing the Markov property (2.12) with time series graph $\mathcal{G}$. To simplify notation, we drop the time indices. We assume that X, Y are connected by a directed path with path nodes $\mathcal{C}_{X \to Y}$ including $X_{t-\tau}$ as defined in Sect. 2.4.3. We denote those parents of Y that are in the path nodes as $\mathcal{P}^{\mathcal{C}}_Y = \mathcal{P}_Y \cap \mathcal{C}_{X \to Y}$ and correspondingly for other path nodes and assume the following dependencies:*

$$\begin{aligned} X &= g_X(\mathcal{P}_X) + \eta^X \\ Y &= f_Y(\mathcal{P}^{\mathcal{C}}_Y) + g_Y(\mathcal{P}_Y \setminus \mathcal{P}^{\mathcal{C}}_Y) + \eta^Y, \end{aligned} \tag{5.45}$$

where f_Y is linear and g_Y arbitrary. Further, for all path nodes $Z^{(i)}$ we assume the dependencies

$$Z^{(i)} = f_i(\mathcal{P}_i^{\mathcal{C}}) + g_i(\mathcal{P}_i \setminus \mathcal{P}_i^{\mathcal{C}}) + \eta^i \quad \forall \; Z^{(i)} \in \mathcal{C}_{X\to Y} \setminus \{X_{t-\tau}\}, \tag{5.46}$$

where the f_i are again linear and the g_i arbitrary. Then MITP given by Eq. (3.53) reduces to a mutual information

$$I^{\text{MITP}}_{X\to Y} = I(\eta^Y + f(\eta^X, \cup_i \eta^i); \eta^X), \tag{5.47}$$

where f is a linear function and $\cup_i \eta^i$ denotes the innovation terms of all path nodes in $\mathcal{C}_{X\to Y} \setminus \{X_{t-\tau}\}$.

The proof is given in Appendix A.4.2.

Momentary interaction information (MII)

Since momentary interaction information (MII) is the difference between MITP and the MITP conditioned on one of the path nodes (excluding $X_{t-\tau}$), the dependencies follow from the above theorems.

Theorem 5.5 (Coupling strength autonomy for MII)**.** *Using the same assumptions as for Theorem 5.4, the momentary interaction information $\mathcal{I}^{\text{MII}}_{X\to Y|Z}(\tau, \tau_Z)$ between $X_{t-\tau}$, Y_t and an intermediate process $Z_{t-\tau_Z} \in \mathcal{C}_{X\to Y} \setminus \{X_{t-\tau}\}$ reduces to*

$$\mathcal{I}\left(\eta^X; \eta^Y + f(\eta^X, \cup_i \eta^i); \eta^Z + f_Z(\eta^X, \cup_i \eta^i \setminus \{\eta^Z\})\right), \tag{5.48}$$

for linear functions f, f_Z.

The proof is given in Appendix A.4.2. For the special case of a causal triple as shown in Fig. 5.5 this further reduces to

$$\mathcal{I}\left(\eta^X; \eta^Y + (c + ab)\eta^X + b\eta^Z; \eta^Z + a\eta^X\right). \tag{5.49}$$

MIT, MITP and MII somewhat disentangle the coupling structure, which is exactly the coupling strength autonomy that makes these measures well-interpretable as measures that solely depend on the "coupling mechanism" between $X_{t-\tau}$ and Y_t (and possibly intermediate processes), autonomous of other external processes. In statistics a measure with such an 'invariance' property is called an *ancillary statistic* (Ghosh et al., 2010). More precisely, an ancillary statistic is a statistic whose sampling distribution does not depend on the parameters of the model, in our case the coefficients of the process equations. One such possible misleading input "filtered out" by MIT is autocorrelation, or, more generally, autodependency as has been shown in the studies of significance in Sect. 4.3.3 and will be further demonstrated in numerical experiments in Sect. 5.4 and the application to climatological data in

Chapter 6. In this way, MIT, MITP, and MII can be framed under the paradigm of conditional inference (Reid, 1995) discussed in Sect. 3.6 which aims at eliminating nuisance parameters in order to more accurately infer a causal interaction strength.

Drawing a possibly far-fetched analogy, the goal in independent component analysis (e.g., Stögbauer et al. (2004)) is to decompose a multidimensional time series into maximally independent components that might be attributable to distinct subprocesses. In the linear case, principal component analysis is widely used to reduce dimensions of large gridded time series datasets and the estimated components even constitute the definition of certain climate indices (Von Storch and Zwiers, 2002). Now the coupling strength autonomy property establishes that MIT provides *maximally independent* measures of interaction among the different subcomponents in the following sense: The coupling strength theorem for MIT implies that under the conditions (5.35 ,5.36, 5.37) the MIT is independent of other coefficients belonging to other links. If this holds for all coupling strengths of all links in the process, then the MITs are independent in a functional sense. Note, however, that all coupling strengths of links emanating from the same process X will depend on the source variance of η^X. MI between X and Y, on the other hand, is strongly depends on the auto-MITs of X and Y and other external processes. The advantage of MIT is, thus, that the different contributions to an interaction between X and Y can be evaluated. In the introduction we termed these the internal strength of X which can be quantified by the auto-MIT $I^{\text{MIT}}_{X\to X}$, further $I^{\text{MIT}}_{X\to Y}$ quantifies the coupling mechanism's strength, and $I^{\text{MIT}}_{Y\to Y}$ the susceptibility of Y. These terms will be interpreted further in the physical interpretations in Sect. 5.5.

5.4. Numerical comparison of dependency measures

In the following, we compare MI, TE, MIT and related measures numerically to investigate the properties of generality and coupling strength autonomy for a general class of nonlinear discrete-time stochastic multivariate processes (Hastie and Tibshirani, 1986):

$$\begin{aligned}
Z_t &= a_Z Z_{t-1} + \eta^Z_t \\
X_t &= a_X X_{t-1} + c_{ZX}\ g(Z_{t-1}) + \eta^X_t \\
Y_t &= a_Y Y_{t-1} + c_{WY}\ g(W_{t-1}) + c_{XY}\ f(X_{t-2}) + \eta^Y_t \\
W_t &= a_W W_{t-1} + \eta^W_t
\end{aligned} \tag{5.50}$$

with independent Gaussian white noise processes η^i_t with all variances $\sigma^2_\cdot = 1$. The corresponding time series graph is depicted in Fig. 3.5(b). We estimate the various coupling measures for different c_{XY} and fixed $a_Z = a_W = 0.5$ and vary the input coefficients

$$\begin{aligned}
a_X = c_{ZX} &\in \{0.0,\, 0.1,\, \ldots,\, 0.8\} \\
a_Y = c_{WY} &\in \{0.0,\, 0.1,\, \ldots,\, 0.8\}
\end{aligned}$$

and functional dependencies of inputs

$$\begin{aligned}
\text{linear} \quad & g(x) = x, \\
\text{squared} \quad & g(x) = 0.3 \cdot x^2, \\
\text{stochastic} \quad & g(x) = 2x\varepsilon_t \quad \text{with uniform i.i.d. } \varepsilon_t \in [0,1], \\
\text{exponential} \quad & g(x) = 0.3 \cdot 2^x, \\
\text{sinusoidal} \quad & g(x) = \sin 4x.
\end{aligned}$$

Here we depict results for linear $f(x) = x$ such that the multivariate process suffices all three conditions, and nonlinear dependency $f(x) = x^2$. The ensemble E then is defined by all combinations of input coefficients and functional forms, each combination run with 120 trials. The CMIs are estimated using the nearest-neighbor estimator as discussed in Sect. 4.2 with parameter $k = 1$ (small values of k lead to a lower estimation bias but higher variance).

5.4.1. Coupling strength autonomy

In the top panel of Fig. 5.7(a), we plot the ensemble average $\left\langle \hat{I} \right\rangle_E$ for fixed $c_{XY} = 0.6$ for the following measures with $\tau = 2$: MI (Eq. (3.37), gray with dotted line), ITY (Eq. (3.45), green with dash-dotted line), ITX (Eq. (3.51), blue with dashed line) and MIT (Eq. (3.48), red with solid line). The parents are shown in Fig. 3.5(b).

MIT is largely invariant to changes of the remaining coefficients and $g(x)$ and approximately attains the analytical value for zero input coefficients (given by Eq. (5.16) for $c_{XY} = 0.6$ and $\sigma_X^2 = \sigma_Y^2 = 1$): $I \approx 0.15$. This implies that the MIT of the coupling link is autonomous of the MITs corresponding to the input links $Z \to X$ for $Z \in \mathcal{P}_X$ and $W \to Y$ for $W \in \mathcal{P}_Y \backslash \{X\}$ which scale with these coefficients. Note, however, that all coupling strengths of links emanating from the same process will depend on its variance $\sigma_\cdot^2$ like in Eq. (5.16). Further, MI is mostly larger, but can also be smaller than MIT, which can be explained with the entropy diagram in Fig. 3.5(b): larger MIs occur if the entropy is increased due to a larger input of $H(\mathcal{P}_X)$ and smaller MIs occur if the relative shared part of $H(X)$ in $H(Y)$ decreases due to a larger input of $H(\mathcal{P}_Y)$. For zero inputs, MI approaches the analytical value $I \approx 0.15$ where all four measures converge to. ITY can at least exclude input to Y and ITX can exclude input to X. Note, however, that the dependence of ITX and ITY on the input coefficients can be different in other models (in Sect. 5.2.1 we found that ITY can even still depend on the auto-dependency of Y). The average of ITX (ITY) is always smaller (larger) equal than MIT confirming the inequality Eq. (5.38).

In the bottom panel of Fig. 5.7(a), we compare MIT (red with solid line) to TE according to Eq. (3.39) truncated at $\tau_{\max} = 4$ (gray with dotted line), the CMI $I^{\text{LINK}}_{X \to Y}$ defining links in the time series graph according to Eq. (3.43) truncated at $\tau_{\max} = 4$ (green with dash-dotted line), and DTE according to Eq. (3.42) with $\tau^\star = 3$ (blue with dashed line). TE and LINK have a much larger estimation dimension of 17 (as much as 25 for $\tau_{\max} = 6$) compared to 6 for MIT and between 5 and 12 for the

summands of DTE. As analyzed in Sect. 4.2.3, the higher dimensional estimation of CMI is strongly biased and here this leads to a negative relative bias in TE of about 50% for the analytically known value for zero input coefficients $I \approx 0.15$. DTE, on the other hand, is much less biased. Apart from this bias, TE and DTE scale similarly with the input coefficients. LINK is dependent on a_X as we expect from our analytical considerations (Eq. (5.12)). MIT shows some slight dependence for strong inputs due to estimation problems for short samples, but overall, also numerically we demonstrate here that only MIT fulfills the proposed property of coupling strength autonomy.

5.4.2. Multivariate equitability

In Fig. 5.7(b) we show the whole densities of E of all measures for *different* coupling coefficients c_{XY}. The aim of this experiment is to measure how well the measures can distinguish the coupling strength for different c_{XY} as demanded by the property of multivariate equitability. The dashed lines show the densities of the ensemble for $a_X = c_{ZX} = a_Y = c_{WY} = 0$, i.e., if both X and Y are independent of their parents.

As we now already expect, MI takes a whole range of values for the same c_{XY}. ITY is broadly peaked towards higher I values and ITX towards lower values, confirming the inequality Eq. (5.38). Note that this relation holds only on average. Only with MIT the different coupling coefficients c_{XY} can be well distinguished. DTE tends to slightly higher values for larger autodependencies within X as expected from our analytical results in Sect. 5.2.2. Additionally, the variance of the DTE estimate is higher because each summand's variance adds up to the total variance of the DTE estimate. The remaining four plots demonstrate that TE and LINK strongly suffer from the negative bias associated with high dimensional estimation depending on the chosen $\tau_{\max}$ discussed in Sect. 4.2.3. TEs or LINKs estimated with different $\tau_{\max}$ can, therefore, not be compared with each other.

For the 'unperturbed' case of zero inputs, the ensemble distributions of MI (dashed lines in Fig. 5.7(b)) are – as expected – similar to the one for MIT with "conditioned-out" inputs (solid lines) apart from a small bias and smaller variance related to slightly higher dimensional estimation. For conditionally independent variables ($c_{XY} = 0$, red lines), all measures have small bias (however, as studied in Sect. 4.2.3 for stronger driving the bias increases). It may seem that apart from the bias, at least the variance is much smaller for the high dimensional measures TE and LINK, but the relative variance $\left\langle \hat{I}^2 \right\rangle / \left\langle \hat{I} \right\rangle$ actually increases leading to a worsened distinguishability.

Regarding equitability, a desired property of a coupling measure would be that it scales linearly with the coupling parameter c_{XY} like the partial correlation approximately in the Gaussian case. As can be seen from the analytical derivations and the numerical example in Fig. 5.7(b), MIT scales $\propto \ln(1 + c_{XY} \cdots)$ for Gaussian dependencies, but a linear scaling in this case can be attained by the transformation $I \to \sqrt{1 - e^{-2I}} \in [0, 1]$ as discussed in Sect. 4.2.5. For more complex dependencies improved estimators that are more adapted to the distributions might help.

In Fig. 5.8 we show results of our numerical experiments for the model class of Eq. (5.50) with a nonlinear dependency $f(x) = x^2$ of the link "$X_{t-2} \to Y_t$" using the same ensemble setup E as before. As discussed in Sect. 5.2.5, then the source process $\eta^X_{t-\tau}$ mixes with its parents and it does not make sense to attribute the coupling strength to one single coefficient. As a result, the average of MIT in Fig. 5.8(a) tends to larger values for increased $a_X = c_{ZX}$, thus the inputs are not entirely "filtered out". Still, MIT is much less affected than MI. Regarding the inequality relation Eq. (5.38), a nonlinear dependency does not affect at least the right-hand side $I^{\mathrm{MIT}}_{X\to Y}(\tau) \leq I^{\mathrm{ITY}}_{X\to Y}(\tau)$ as demonstrated in Fig. 5.8(a) and (b). Although the left-hand side of the inequality relation $I^{\mathrm{ITX}}_{X\to Y}(\tau) \leq I^{\mathrm{MIT}}_{X\to Y}(\tau)$ should hold under the same general Markov property and the "no sidepath"-condition, it seems to be violated for large $a_X = c_{ZX}$ (and small $a_Y = c_{WY}$). This could be related to highly skewed distributions for nonlinear $f(x)$.

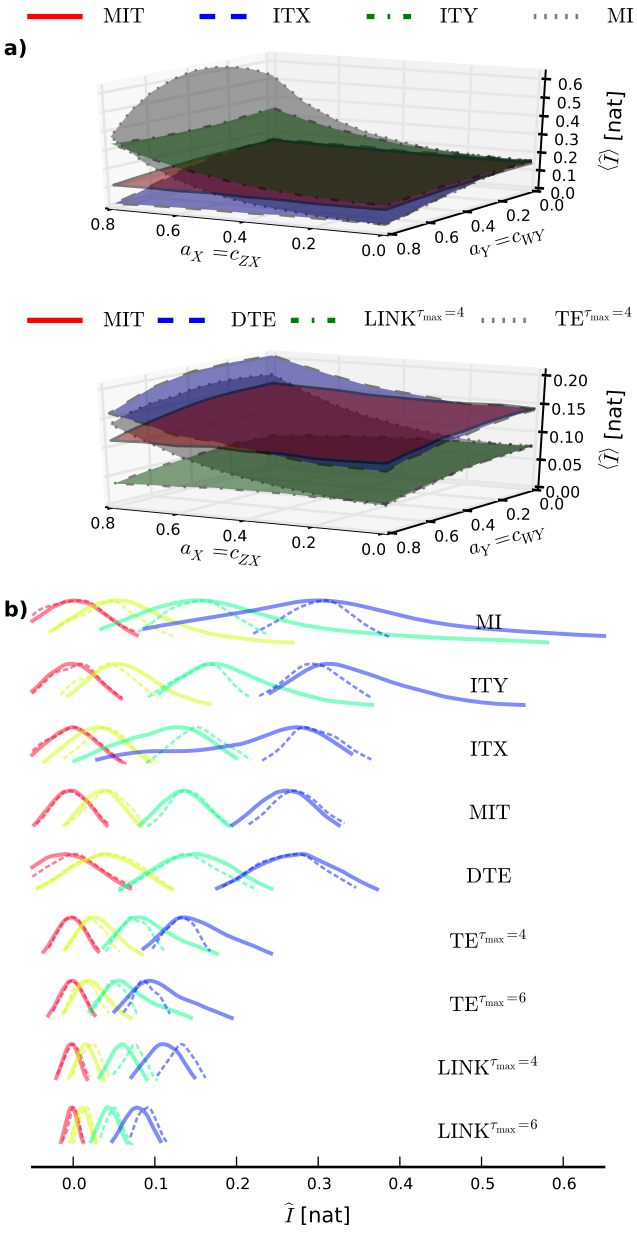


Figure 5.7.: Numerical experiments with the model Eq. (5.50) using time series length $T = 1000$. (a) Ensemble average $\langle I \rangle_E$ for fixed $c_{XY} = 0.6$ for all measures as specified in the main text. (b) Ensemble densities of all measures for different coupling coefficients $c_{XY} = 0.0, 0.3, 0.6, 0.9$ (from left to right red, yellow, green and blue solid lines). The densities are estimated using Gaussian kernel smoothing according to Scott's rule, showing only the 90% most probable ensemble members.

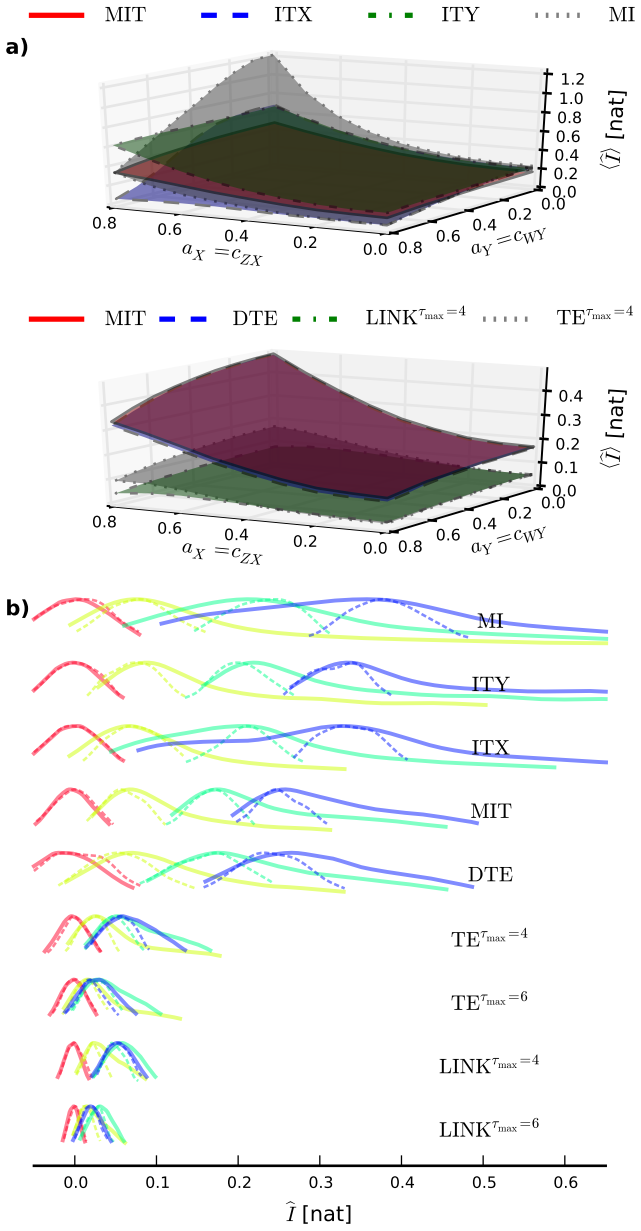


Figure 5.8.: Numerical experiments with the model Eq. (5.50) with setup as before but for squared dependency $f(x) = x^2$.

5.5. Physical interpretation and discussion

In the following sections, we relate our information theoretic approach to communication theory, thermodynamics and discuss it in the geophysical context following Runge et al. (2012b); Runge et al. (2014). Further, we study how the path-based measures can complement measures from complex network theory.

5.5.1. Communication theory

The form of Eq. (5.44) is reminiscent of the Shannon-Hartley theorem in communication theory (Cover and Thomas, 2006). Communication theory is closely linked to information theory (Shannon, 1948). It deals with the problem of the limits of efficient transmission of messages over a communication channel. The communication channel capacity C is given by the maximum MI over *all possible* input sources: $C = \max_{\{P(X)\}} I(X;Y)$. The Shannon-Hartley theorem for channels subject to additive Gaussian white noise of power N then reads

$$C = B \log \left(1 + \frac{S}{N}\right) \tag{5.51}$$

with bandwidth B and signal-to-noise ratio S/N. Maximizing over all possible input sources gives a notion of the *potential* influence transmittable, while MI or MIT measure the *observed* influence (Janzing et al., 2013). That is, the difference to our measure of coupling strength is that we cannot manipulate the input sources and thus cannot measure the channel capacity alone. Some attempt in this direction can be found in Permuter and Naiss (2010). We also expressed the various other CMIs occurring above in the form of the Shannon-Hartley theorem, for example, in Eq. (5.15) $c_{XY}^2 \sigma_X^2$ is the signal strength and $c_{WY}^2 \sigma_W^2 + \sigma_Y^2$ is the noise strength. For MIT the signal-to-noise ratio S/N corresponds to $c_{XY}^2 \sigma_X^2 / \sigma_Y^2$ (Eq. (5.44)), that is, we measure the sources variance σ_X^2, modified by the channel parameter c_{XY}^2 as the signal in relation to the noise coming from the target alone.

Comparison to causal strength defined in Janzing et al. (2013) In Janzing et al. (2013), another information-theoretic approach to quantify causal strength, based on a different set of postulates, is discussed. Their idea is inspired by communication theory in that they assess the importance of a link by the impact of corrupting it, i.e., cutting the wire. The "open ends" are then fed with their marginal distributions which emulates the idea that an attacker blurs her attack with the only distribution she can see. For the case of only one cut link the same idea is pursued from an interventionalist perspective in Ay and Polani (2008). Janzing et al. (2013) give an example of causal strength in the chain $Z \to X \to Y$. For the (pathological) case that $Z \to X$ is a copy operation, their measure of strength between X and Y is non-zero. MIT, on the other hand, would be zero, because all information between X and Y is already contained in Z. That is, no information is generated at all in X, i.e., its source entropy is zero. We believe that this is an unrealistic assumption for

complex time dependent systems and even if this case occurs, we would not interpret the link $X \to Y$ as a physical mechanism, but only the indirect link $Z \to X \to Y$ and assess a coupling strength using MITP for this chain mechanism. Generally, each approach has its "intuitive" justification and the practitioner has to decide which one is best for his application.

5.5.2. Thermodynamics

While the long standing discussion between information theory and the foundations of statistical mechanics (Jaynes, 1957; Crutchfield and Shalizi, 1999; Allahverdyan et al., 2009) is beyond the scope of this work, we will give a modest approach to interpret MIT thermodynamically here.

Following the derivation for the case of transfer entropy in Prokopenko et al. (2013), we start by considering the specialized Boltzmann's principle as used in Einstein (1905),

$$S - S_0 = k_B \log W_r \tag{5.52}$$

where S_0 is the entropy of an equilibrium and S of a non-equilibrium state, k_B is the Boltzmann constant and W_r is the probability of a transition between the two states, more precisely, the ratio of W and W_0 that account for the numbers of microstates in the macrostates with S and S_0, respectively.

Without loss of generality, this change in entropy $\Delta S = S - S_0$ can be decomposed into the contributions due to internal and external interactions (a similar heuristic approach has been used in Liang (2013))

$$\Delta S = \Delta S_{ext} + \sigma. \tag{5.53}$$

To relate this to the MIT framework of couplings, we now assume the equilibrium state of Y to be y_t *in the context of/together with* its parents $\mathcal{P}_{Y_t} \backslash \{x_{t-\tau}\}$ and the parents of X at lag τ, $\mathcal{P}_{X_{t-\tau}}$, subsumed as z_t, and attribute the external entropy production to X at lag τ. The internal entropy change we associate with the source entropy of Y.

To this end, we consider the reversible transition in system Y from state y_t to y_{t+1} in the context of Z which corresponds to some number W_r' such that the entropy change $\Delta S = S(y_{t+1}) - S(y_t) = k_B \log W_r'$ and hence

$$p(y_{t+1}|y_t, z_t) = \frac{1}{Z'} e^{(S(y_{t+1}) - S(y_t))/k_B}, \tag{5.54}$$

where Z' is some normalization constant that depends on y_t and z_t.

Secondly, we relate the irreversible transition from y_t to y_{t+1} in the context of z_t *and* $x_{t-\tau}$ to some number W_r'' such that the source entropy change $\sigma = S(y_{t+1}) - S(y_t) =$

$k_B \log W'_r$ and hence

$$p(y_{t+1}|y_t, z_t, x_{t-\tau}) = \frac{1}{Z''} e^{\sigma/k_B}, \tag{5.55}$$

where Z'' is some normalization constant that depends on y_t and z_t.

Formulating MIT not as an average, but as an information rate

$$i^{\mathrm{MIT}}_{X\to Y}(\tau) = h(y_{t+1}|y_t, z_t) - h(y_{t+1}|y_t, z_t, x_{t-\tau}), \tag{5.56}$$

we can now link the entropy from external interactions to MIT via (disregarding unimportant constants)

$$\begin{aligned} h(y_{t+1}|y_t, z_t) &= -\ln p(y_{t+1}|y_t, z_t) && (5.57)\\ &= \ln Z' - \tfrac{1}{k_B}(S(y_{t+1}) - S(y_t)) && (5.58)\\ h(y_{t+1}|y_t, z_t, x_{t-\tau}) &= -\ln p(y_{t+1}|y_t, z_t, x_{t-\tau}) && (5.59)\\ &= \ln Z'' - \tfrac{1}{k_B}\sigma, && (5.60) \end{aligned}$$

and arrive at

$$i^{\mathrm{MIT}}_{X\to Y}(\tau) = \ln \tfrac{Z'}{Z''} + \tfrac{1}{k_B}(\sigma - (S(y_{t+1}) - S(y_t))). \tag{5.61}$$

For small fluctuations near equilibrium, i.e., – statistically speaking – for stationary time series, we have $Z' = Z''$ and thus

$$i^{\mathrm{MIT}}_{X\to Y}(\tau) = -\tfrac{1}{k_B}(\Delta S_{ext}). \tag{5.62}$$

The MIT rate can, therefore, be interpreted as being proportional to the external entropy production. In the way we put it, the MIT rate measures the difference in entropy rates between the (by us defined) reversible process and the irreversible process affected by another source X. Note that the information rate can be negative while the average MIT is always non-negative. This derivation is clearly not a very satisfactory one because we have defined the equilibrium state in an *ad hoc* manner including the past of the driving system X. For the case of transfer entropy the equilibrium state is solely given by Y (Prokopenko et al., 2013). MIT is a complex measure for such a thermodynamic interpretation.

5.5.3. Geophysics

There are also observational examples that agree with the analytical findings regarding the delays obtained from the cross correlation lag function in Section 5.2.1. To name just two, in the example from Gu and Adler (2011) mentioned in the introductional Section 2.2.2, El Niño-Southern Oscillation (ENSO) was found to influence land precipitation with much shorter lags compared to land temperatures (their Figs. 4(c,d)). In the light of our analysis this finding can be interpreted differently: The coupling

delay of the mechanism of ENSO's influence on temperature and precipitation might be the same and just the precipitation has a much lower autocorrelation as often is the case for precipitation data. Also in Huang et al. (2011) the correlations of meteorological variables on malaria are found to be much weaker after *prewhitening* the time series. Prewhitening refers to the procedure to fit and remove an first-order autoregressive model model from the time series which obviously decreases serial correlation.

How can these results be interpreted physically? And what do these results mean for the interpretability of correlation as a measure of the delay and 'link strength' of a mechanism? We will try to provide an intuition by considering the following very simple bivariate stochastic climate model (Frankignoul and Hasselmann, 1977; Von Storch and Zwiers, 2002):

$$\begin{aligned} \tfrac{d}{dt}X(t) &= -\tilde{a}X(t) + \tilde{\varepsilon}^X(t) \\ \tfrac{d}{dt}Y(t) &= -\tilde{b}Y(t) + \tilde{c}X(t) + \tilde{\varepsilon}^Y(t). \end{aligned} \tag{5.63}$$

As shown in Fig. 5.9, this Ornstein-Uhlenbeck system can be understood as a particle or 'climatic variable' Y that fluctuates around an equilibrium state in a parabolic potential $V(Y) = \frac{1}{2}\tilde{b}Y^2$ from which an internal force $-\frac{\partial}{\partial Y}V(Y) = -\tilde{b}Y$ results. This system is driven by another 'climatic variable' X and random noise $\tilde{\varepsilon}^Y$ with a very short decorrelation time such that it can be approximated by white noise with variance $\tilde{\sigma_Y}^2$. The variable X is considered to fluctuate in its own potential, driven only by white noise independent of the other noise term. Now small coefficients $\tilde{a}$ and $\tilde{b}$ correspond to shallow potential wells and vice versa, and $\tilde{c}$ reflects the strength of the mechanism by which variability in X influences the particle position Y. If Eq. (5.63) is discretized and Taylor expanded for small time steps $\mathcal{O}(\Delta t)$, one arrives at the same bivariate autoregressive model Eq. (5.1) studied before for $\sigma_{XY} = 0$ and with the substitutions (Brockwell and Davis, 2002)

$$\begin{aligned} &a \to 1 - \tilde{a}\Delta t, \quad b \to 1 - \tilde{b}\Delta t, \quad c \to \tilde{c}\Delta t, \\ &\sigma_X^2 \to \tilde{\sigma_X}^2 \Delta t, \quad \sigma_Y^2 \to \tilde{\sigma_Y}^2 \Delta t. \end{aligned} \tag{5.64}$$

Therefore, we can now interpret the behavior of the cross correlation function (which also applies to regressions) in terms of a particle Y. As visualized in Fig. 5.9, a strong autocorrelation in Y (large b) can then be understood as a very shallow potential well (small $\tilde{b}$) which leads to the particle taking large departures from its equilibrium position before slowly coming back giving rise to a strong persistence in the time series. A shallow potential well also renders the particle Y more susceptible to external fluctuations. If also the particle X is immersed in a shallow potential well, a more persistent external force is exerted on Y. Thus, for large autocorrelations these two effects act together leading to a large covariation of X and Y slowly decaying back to their equilibria, which implies that even if the coupling strength c is small, X has a large effect on Y and consequently there is a larger cross correlation. Further, for increasing b the delay is shifted towards larger lags due to the strong inertia of Y.

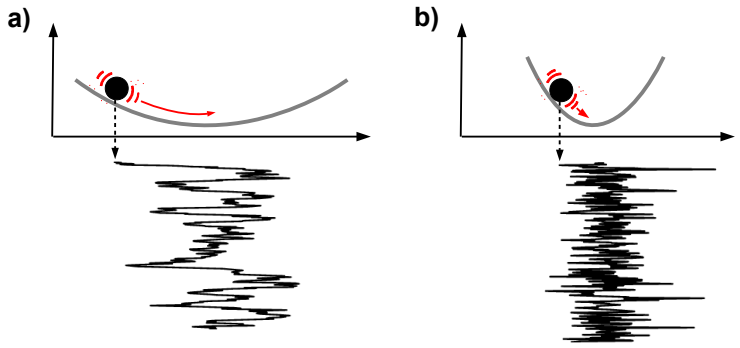


Figure 5.9.: Physical picture of persistence or autocorrelation via a model of a particle in a potential well subject to stochastic forcing: (a) Shallow potential well leading to large autocorrelation and (b) narrower well leading to a weaker autocorrelation.

But where does the 'shallow potential well' come from in the example of surface temperatures in the tropics and where does the 'narrower potential well' come from in Europe shown in our motivational example in Sect. 2.2.2? Geophysically, one reason for more inertia in the tropical surface temperatures above the ocean is the higher specific heat capacity of the ocean leading to dampened temperature fluctuations. Accordingly, the almost vanishing autocorrelation for the European time series (Fig. 2.1) can be well explained by the 'short-term memory' of the mid-latitude atmosphere. More climatological interpretations will be given in comparison with the measures MIT and ITY in Sect. 6.3.

The important point now is that only the coefficient c reflects the actual factor of the mechanism by which X influences Y. And since this factor is still dependent on the units of the variables, only c normalized by the innovation's variances reflects the actual strength of the mechanism. Further, only the lag occurring in the physical equation reflects the actual delay (e.g., in delay differential equations). Then, again, we stress that the analysis implies that the cross correlation is not even a good first order approximation of the coupling strength and the maximum's lag also not a good indicator of the coupling delay of the mechanism. Rather, the analysis demonstrates their strong sensitivity even on slight deviations in high autocorrelation. But should a measure of coupling strength and delay between X and Y depend on their internal dynamics, here given by the width of their potentials?

In this chapter, we have proposed MIT as a measure that excludes these influences in assessing the strength and lag of a mechanism. In the following Chapter 6, we will give climatic examples that demonstrate how MIT can be used to better understand physical mechanisms because MIT is not as ambiguous and better interpretable than cross correlation or mutual information and also transfer entropy.

5.5.4. Information transfer and complex network theory

As mentioned in the introductory chapter, in the literature of neuroscience (Bullmore and Sporns, 2009; Blinowska and Kaminski, 2013; Simpson et al., 2013) and recently also in climate research (Donges et al., 2009a; Donges, 2012), multivariate datasets are often analyzed using pairwise association measures. Then this association matrix is thresholded (either by some predefined threshold or such that a fixed link density is obtained) and the resulting binary adjacency matrix is studied from a graph-theoretical perspective using statistical network theory (Newman, 2010). In interpreting such networks, it is important to take into account the aspect that the network comes from only pairwise associations. For example, the basic principle of transitivity of correlation leads to a lot of spurious links strongly affecting network measures such as the average path length. Typically, short-path lengths in these networks are related with global efficiency of 'information transfer', e.g., in the brain (Bullmore and Sporns, 2009), but also climate (Tsonis et al., 2008). But in Sect. 6.5 we will analyze a causal network of atmospheric pressure components and see that, for example, even though there are multiple paths between the components of ENSO and the North Atlantic oscillation in the causal network, they are still not even correlated. Bialonski et al. (2010); Hlinka et al. (2012) have shown that even for a set of entirely independent processes a small world topology emerges. Further, the robustness of a system to random error or perturbations is typically associated with a high clustering coefficient. Also this measure can lead to false interpretations if causality is not taken into account. For example, for the true causal relations $X \to Y \to Z$, there are significant correlations between all pairs and the clustering coefficient of the non-causal network would be maximal. In this simple example an 'attack' on node Y in the center certainly disrupts the causal network most because it also destroys the interlink between X and Z. But this is not taken into account if the non-causal network is analyzed which shows that some doubt can be casted on these interpretations. In recent years some studies in neuroscience have also applied linear Granger causality methods (Liao et al., 2011; Deshpande et al., 2011) and bivariate transfer entropy has been applied to climate time series (Hlinka et al., 2013).

With the measures defined in Sections 5.2.3 and 5.2.4, we make an attempt to put the notion of shortest paths in an information-theoretic perspective. MITP is particularly well suited towards such an interpretation. It measures the impact of a momentary perturbation in X on another not necessarily adjacent node Y in the network. Thus, instead of counting shortest paths between X and Y, MITP gives an appropriate measure of how much information is actually transferred. The interaction information MII can then be seen as an alternative to *betweenness centrality* (Newman, 2010; Boccaletti et al., 2006). Betweenness centrality is defined as

$$g(v) = \sum_{i \neq v \neq j} \frac{s_{st}(v)}{s_{st}}, \tag{5.65}$$

where s_{st} is the total number of shortest paths from node i to node j and $s_{st}(v)$ is the number of those paths that pass through v. Recall that MII, given by

$$\mathcal{I}^{\text{MII}}_{X\to Y|Z}(\tau,\tau_Z) = I^{\text{MITP}}_{X\to Y}(\tau) - \underbrace{I(X_{t-\tau};Y_t \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t},\mathcal{P}(\mathcal{C}_{X_{t-\tau}\to Y_t}),Z_{t-\tau_Z})}_{\text{MITP additionally conditioned on } Z_{t-\tau_Z}}, \quad (5.66)$$

quantifies how much an intermediate node Z on a causal path between X and Y changes MITP. A *causal interaction betweenness* can then be defined as the aggregated sum of all interactions that are significantly changed by Z, possibly differentiating between counteracting and enhancing changes. Also the related measure ITP can be used in such a way if not the influence of momentary perturbations are of interest, but *all* information entering the system in X. In our applications to climate data in Sect. 6.5 we will demonstrate how these measures can be used to determine processes important for distributing and mediating information.

5.6. Underlying assumptions and limitations of inferring causal strength

While we have discussed the limitations of inferring causality from observed time series alone already in Chapter 2, here we discuss the underlying assumptions and limitations of assessing the strength of a causal mechanism.

Firstly, the graphical model imposes a discrete description of causal interactions. Regarding the source entropy, we face the problem that if a time-continuous process is sampled at some interval Δs, there is an infinite set of unobserved nodes in between every X_t and X_{t-1} for $X \in \mathbf{X}$ in the time series graph. We will, therefore, not be able to assess the source entropy solely at time t, but only the aggregated information in the interval $[t-\Delta s, t]$.

As discussed in the coupling strength autonomy theorem, not in all cases a coupling strength can be attributed to only one single coefficient. Only if this is the case, i.e., under the conditions (5.35)–(5.36) in Theorem 5.3, MIT can filter out all influences from the parents of X and Y. If the dependency is nonlinear or sidepaths exist, one could use modifications of MIT like $I^{\text{MITP}}_{X\to Y}$ [Eq. (3.53)] or a multivariate CMI like $I^{\text{SITY}}_{X\to Y}$ [Eq. (3.60)] for a more appropriate measure of coupling strength. Note that even though for full coupling strength autonomy the link "$X_{t-\tau} \to Y_t$" needs to be linear, the remaining dependencies can still be nonlinear and the source processes can have arbitrary distributions. The process can, therefore, not easily be estimated using model-based regressions.

5.7. Summary

Summarizing, in this chapter we have analytically and numerically studied the measures introduced in Chapter 3 to develop a statistical and physical understanding.

We have analyzed the lagged cross correlation function and regressions as common measures in climate science to identify interaction mechanisms between climatological processes, in particular to assess possible time delays of a mechanism and as a measure to quantify the strength of the link mediated by the mechanism. We have investigated how justified such an approach is in the presence of large autocorrelations which typically occur in tropical temperature time series. An analytical study of a simple autoregressive model suggests that univariate regression coefficients and the cross correlation lag function's maximum value and lag are very sensitive to even slight changes in high autocorrelation. Using the picture of a particle in a shallow potential, we also give a physical explanation for this effect and come to the conclusion that cross correlation and univariate regressions are quite ambiguously influenced by internal dynamics with strong inertia (e.g., a large oceanic heat capacity) and misguide an estimate of a physical coupling strength. The same conclusions also hold for lag functions of other measures like mutual information if the effect of autodependencies is not conditioned out.

Further, we have shown analytically and numerically that the commonly used measures mutual information and transfer entropy can be rather unintuitive as measures of coupling strength. The novel measures based on momentary information overcome this limitation due to a property which we call coupling strength autonomy. It allows for a well interpretable coupling strength of links as well as paths reminiscent of the idea of independent component analysis to provide maximally independent measures of an interaction mechanism. As we prove analytically and numerically, the coupling strength autonomy property allows to entirely decompose the coupling mechanism of a link or along paths for linear interactions, where the external drivers can still be nonlinear. For nonlinear interactions the external drivers cannot be entirely excluded, but still MIT can be well interpreted information-theoretically and is much less affected than mutual information. For such nonlinear interactions we suggest modifications of MIT as more appropriate measures (Sect. 3.5.3). A further advantage compared to transfer entropy is that MIT and related measures are practically computable without the need for arbitrary truncations. Here, TE has the further disadvantage that in order to include larger coupling delays, the bias strongly increases as shown in Sect. 5.4.

In Sect. 5.5, we have related these statistical and information-theoretical results to a physical interpretation in communication theory, thermodynamics and geophysics, where we found that MIT allows to separate the influence of, e.g., the oceanic heat capacity from an assessment of the strength of a coupling mechanism. In Sect. 5.5.4, we have outlined how the novel measures can help to complement measures from complex network theory, such as betweenness centrality, in a way that tries to overcome possible pitfalls if these networks are inferred from only pairwise associations.

Our two-step approach promises to not only extract the causal direct (rather than the indirect) connectivity among processes, but also to assess a meaningful coupling strength, that – together with the coupling delay – assists a physical interpretation. In the next chapter we will demonstrate this interpretation in climate applications.

Part II.

Applications

In this second part, the novel methods are applied to test and generate hypotheses on causal interactions in climate time series covering the 20th century up to the present, in particular teleconnections of the El Niño-Southern Oscillation (ENSO) system. Further, in an exploratory way, a global surface pressure dataset is analyzed to identify key processes that drive and govern interactions in the atmosphere. Finally, it is shown how quantifying interactions can be used to determine possible structural changes, and as optimal predictors, here applied to the prediction of ENSO.

Chapter 6.

Climate interactions

Each system, the body, the brain, the climate, is a universe.

Plamen Ivanov (pers. comm.)

6.1. Introduction – the complex system Earth

Understanding the complex system Earth poses a great challenge for humankind. In the climate system, the interactions of many subsystems from the atmosphere, hydrosphere, cryosphere, and biosphere to human interference with this system are a main focus of today's research, motivated not only by scientific curiosity, but by the urgent need to better understand anthropogenic climate change (Solomon, 2007). In this chapter, we will apply the apparatus of methods developed in this thesis to demonstrate how these analyses can help provide a physical understanding of these complex interactions.

After a concise example of causal interactions in daily sea-level pressure time series over Europe (Sect. 6.2), where we exemplify the causal algorithm and decomposed transfer entropy, in Sect. 6.3 we analyze one of the most important processes in global climate, the El Niño-Southern Oscillation (ENSO). ENSO has a particular political dimension and relevance through its multiple relations to natural disasters such as droughts and floods from South America and Africa to Southeast Asia and Australia (Philander, 1990; Jin, 1997; Cane, 2005). We study ENSO's teleconnections and the main feedback mechanism generating ENSO, the Walker Circulation in Sect. 6.4. For the influence of ENSO on the northern tropical Atlantic, we detect a short lag of one month for this coupling mechanism, while previous studies using the maximum of the cross correlation lag function found lags of 3 to 6 months. Further, our purely statistical analysis confirms the circular causal loop of the Walker circulation. In these analyses we also find novel interpretations of the strengths of these mechanisms. These examples – involving only few processes – shall illustrate the use of our approach to test specific hypotheses on the data and quantify coupling mechanisms.

In a next step in Sect. 6.5, as an example of exploratory data analysis we analyze a globally gridded dataset of sea-level pressure time series. After a dimension reduction yielding distinct components that represent many well-known climatological subprocesses, we analyze their causal interactions drawing on our apparatus from

link-based to path-based measures. Here, we find that ENSO's effect on causally adjacent nodes is large, but comparable to other processes in the tropics. ENSO's outstanding global impact is manifest in its strong effect on non-adjacent processes in the causal network as we infer with path-based measures of information transfer. With MITP and MII, that quantify how much processes impact on the interactions between other processes, we find that next to ENSO and the tropical Atlantic, a process in the East Indian Ocean plays a major role as a process with high causal interaction betweenness (Sect. 5.5.4). Finally, in Sect. 6.6 we give a sliding window analysis of the important impact of ENSO on the Indian Summer Monsoon, where we show that MIT could be used as a proxy for the possible structural change in the weakening coupling between these two processes in recent decades.

These applications of causal approaches to climate data are among the first pioneering examples after the application of graphical models in climate research was recently suggested in Ebert-Uphoff and Deng (2012b); Ebert-Uphoff and Deng (2012a). The short example in the next Sect. 6.2 and the Sect. 6.3 on ENSO's teleconnections and Sect. 6.4 on the Walker circulation summarize and expand results published in Runge et al. (2012a); Runge et al. (2014), while the remaining sections are novel contributions of this thesis.

6.2. Interactions in sea-level pressure over Europe

To illustrate the causal inference algorithm and decomposed transfer entropy (DTE) introduced in Sect. 3.4.2, following Runge et al. (2012a) here we analyze a climatological dataset of daily mean sea level pressure anomalies in the winter months of 1997–2003 (Ansell et al., 2010) at four locations in Eastern Europe indicated on the map in Fig. 6.1. First, we give a statistical analysis and then provide a climatological interpretation demonstrating that our causal picture agrees well with the dynamical processes in this area.

Figure 6.1 shows three iteration steps and the lag functions for the first, corresponding to the MI lag functions, and last step, corresponding to the CMI used in the PC algorithm, i.e., ITY as defined in Section 3.45. From MI in step (0.0) one would infer an almost fully connected graph with a broad range of lags. For example, we found a strong $Y \to Z$ 'link' and a 'link' $W \to X$ with a delay of about two days. The iteration using an initial $n_0 = 2$ converges in the third step (2.1). The link $Y \to Z$ is now much weaker (even below our significance threshold), because a lot of the shared entropy is due to the common driver W. Even more apparent, the $W \to X$ link vanishes due to the condition on Y. Note that the contemporaneous links $X - Z$ and $W - Y$ are possibly due to spatial proximity. The only difference between step (2.0) and (2.1) lies in the incorrect link $W_{t-2} \to Z_t$. We also estimated DTE via Eq. (3.42) with τ^* chosen such that $I(X_{t-\tau^*-1}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau^*-1}})$ has declined below significance.

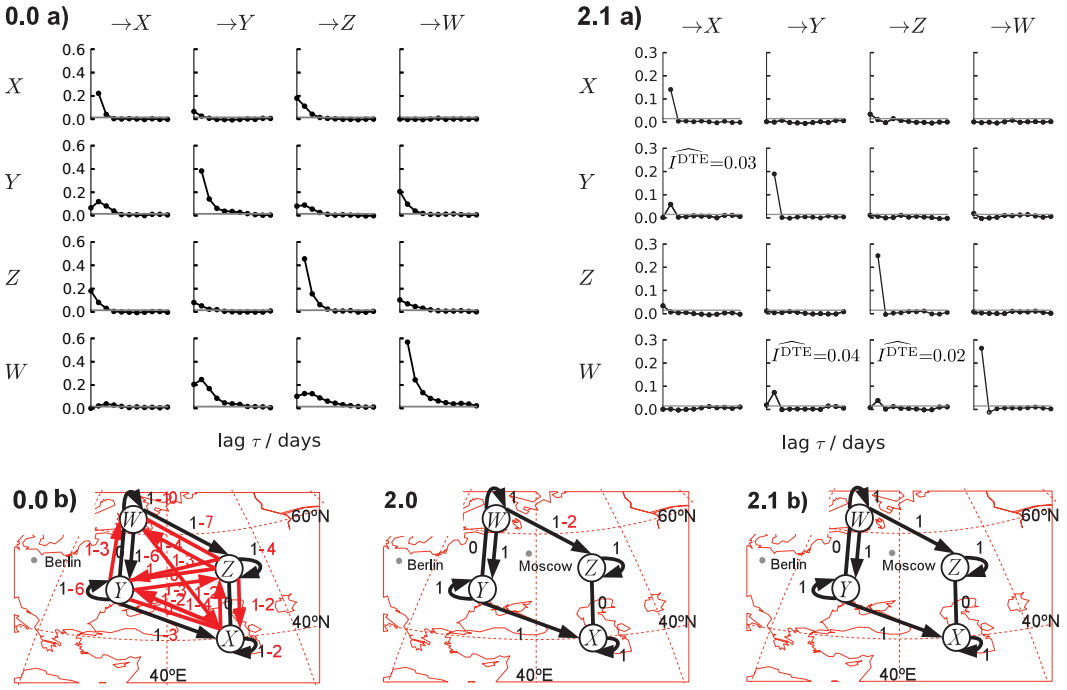


Figure 6.1.: Analysis of daily time series of mean sea level pressure with $T = 1268$ days. The algorithm was run using a threshold $I^* = 0.015$ and $\tau_{\max} = 10$. MI and ITY are estimated with $k = 100$. The label $n.i$ indicates the iteration step. (0.0 a) depicts the MI lag functions, where the gray lines denote the significance threshold, and (0.0 b) the corresponding process graph. With an initial $n_0 = 2$, the next step (2.0) with two conditions is already almost identical to the converged graph in step (2.1). (2.1 a) gives the ITY lag functions, where $\widehat{I^{\mathrm{DTE}}}$ denotes the estimate of DTE via Eq. (3.42) with τ^* chosen such that $I(X_{t-\tau^*-1}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau^*-1}})$ has declined below significance, and (b) again the process graph. 'Incorrect' (i.e., links not found in the converged graph) links and lags are in red.

The DTEs of the links in the south-eastward chain $W \to Y \to X$ are stronger than the eastward link $W \to Z$.

This causal picture of a south-eastward 'flow of entropy' is consistent with the dynamical processes governing the lower and middle atmosphere circulation in the considered area[9]. One usually observes a superposition of westerly winds with traveling extratropical cyclones that traverse the area and whose trajectories are regulated by the aforementioned westerlies (Palmén and Newton, 1969). Consistent with the causal lags of one or two days, these processes act on short daily time scales. We note that this causal structure might change in the high troposphere where the influence of quasi-stationary planetary waves and the Ferrel cell might noticeably modify the above-mentioned causality. This analysis underlines the importance of

[9]The interpretations in this paragraph were mainly contributed by the co-author V. Petoukhov.

inferring coupling delays for physical interpretations and serves as a first step to study more complex systems like ENSO in the next sections.

6.3. ENSO's teleconnections

6.3.1. Statistical analysis

In this section, following Runge et al. (2014), we analyze teleconnections emanating from the El-Niño region towards the Atlantic, South America, and the West Pacific and contrast them with the European West – East link. The comparison between the European and the ENSO – Atlantic link demonstrates effects found in the analytical example from Sect. 5.2.1, where strong autocorrelations shift and inflate the peak of the cross correlation lag function, on real data. Here, we focus on the statistical interpretation, while the results will be discussed climatologically in the next section.

We use the surface air temperature indices Nino3, ATL, WEUR and EEUR analyzed in the motivating example in Sect. 2.2.2. The time series come from the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) reanalysis dataset (Kalnay et al., 1996) and are analyzed for the period 1948–2012 with 780 months. Anomalies, i.e., the subtraction of the seasonal cycle, are taken with respect to the whole period. Nino3 is the time series of the spatial average over the Nino3 region in the East Pacific and ATL is the average over a region in the tropical North Atlantic (all regions are shown on the map in Fig. 6.3). Additionally, we study surface air temperatures in the Eastern (EPAC) from the same dataset and monthly surface pressure anomalies over the western Pacific (WPAC), also from the NCEP/NCAR reanalysis (Kalnay et al., 1996) data set. Further, SSA are monthly precipitation rate anomalies over a region in southern South America (see map in Fig. 6.3) from the Global Precipitation Climatology Project (GPCP) data set in years 1979–2012 (Adler et al., 2003). In this section we use our approach together with the linear partial correlation (allowing for analytical significance tests) to infer the time series graph as well as for MIT and ITY. In all examples we run the algorithm with a maximum time lag of $\tau_{\max} = 15$ months, initial $n_0 = 2$, check 5 different conditioning sets in each step and use a two-tailed significance level of $\alpha = 95\%$. More significance levels are also discussed at the end. Additionally, we show the 5% and 95%-confidence bounds, i.e., the 90% confidence interval. The "±"-values given in the text roughly approximate this interval shown in the figures.

The panels in Fig. 6.2 show the cross correlation and autocorrelation lag functions in light gray. In the same plots we show the values of ITY (blue) and MIT (black), where all non-significant links are marked by gray crosses. For example, the upper right plot in panel (a) shows the lagged cross correlation function $\rho(\mathrm{Nino3}_{t-\tau}; \mathrm{ATL}_t)$ for $\tau \geq 0$ in light gray and the ITY and MIT value at the only significant link "$\mathrm{Nino3}_{t-1} \to \mathrm{ATL}_t$" in blue and black. The estimated parents and neighbors of each variable can be read-off from the ITY values in the columns in this matrix. First,

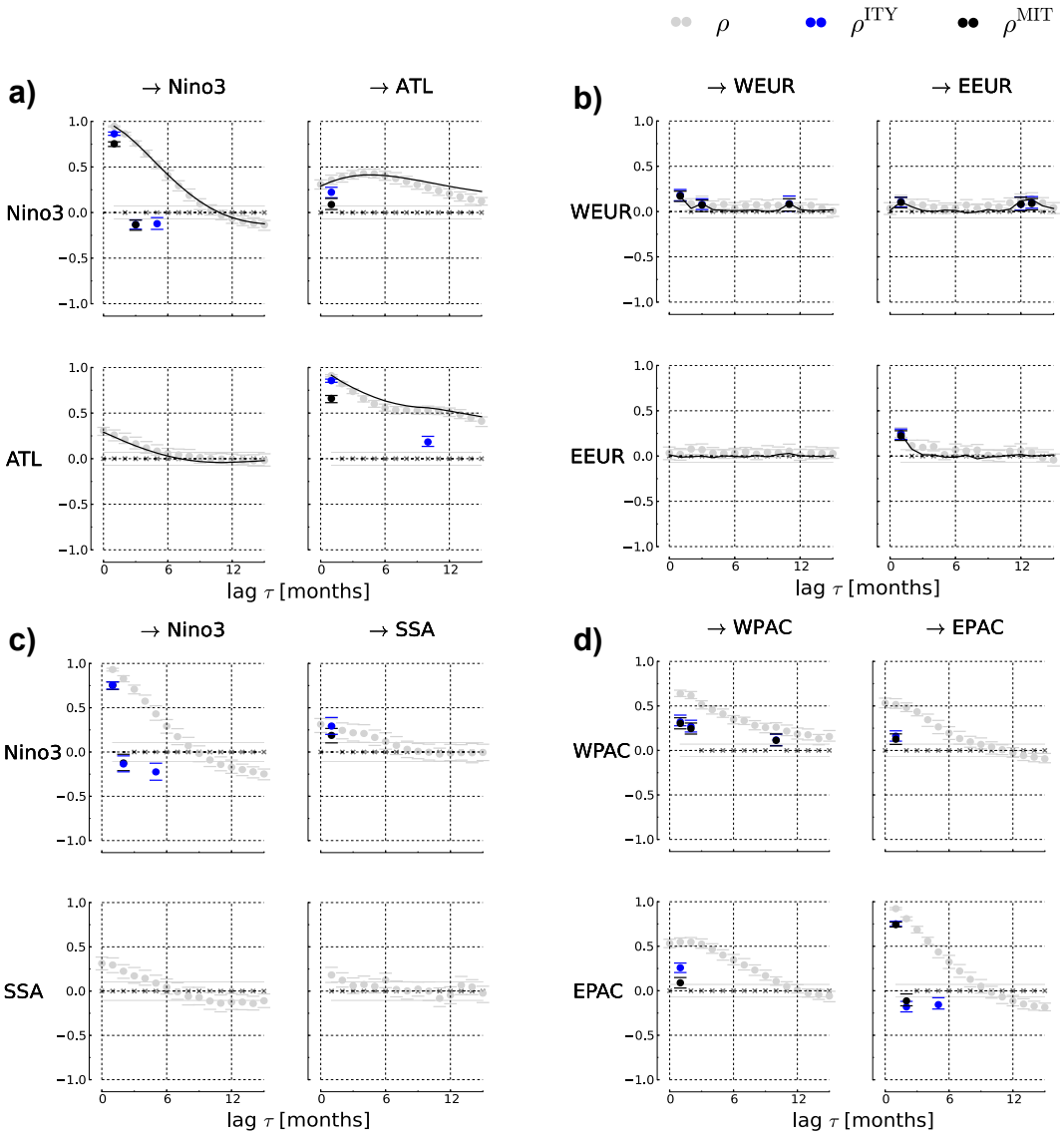


Figure 6.2.: Correlations and partial correlations of four climatic example pairs. The matrix of lag functions in each panel shows the (auto-)correlations (light gray) and the values of ITY (blue) and MIT (black), where non-significant links are marked by gray crosses. The horizontal gray line denotes the two-sided 95%-significance level for the (auto-)correlations. The errorbars mark the 90% confidence interval (in (c) larger due to a smaller sample size). Note that for autocorrelations (on the diagonal) the zero-lag is not drawn. In panels (a) and (b) the solid lines mark the numerically evaluated cross correlations for a Gaussian model fitted to the time series according to Tab. 6.1.

we only compare MIT to the lagged cross correlation and discuss the more subtle differences between ITY and MIT afterward.

Re-examining the motivating example of lag functions in the tropics and Europe, the strength of autocorrelation in Nino3 as measured by MIT in Fig. 6.2(a) is very high and mostly coming from lags at $\tau = 1, 3$ while lags further in the past do not contribute much more for explaining the present. Also ATL has a strong autodependency MIT value at lag 1. In our model example analysis in Sect. 5.2.1, such high autocorrelations resulted in a high and broad peak at a shifted lag in the cross correlation function. Also here most of the 'broad peak links' in Nino3 $\to$ ATL with a maximum at lag 4 are actually non-significant links. Rather, the partial correlation MIT is much smaller than the correlation and significant only at lag 1. For this link MIT is the partial correlation $\rho(\mathrm{Nino3}_{t-1}; \mathrm{ATL}_t | \mathcal{P}_{\mathrm{ATL}_t} \backslash \{\mathrm{Nino3}_{t-1}\}, \mathcal{P}_{\mathrm{Nino3}_{t-1}})$ with parents $\mathcal{P}_{\mathrm{ATL}_t} \backslash \{\mathrm{Nino3}_{t-1}\} = \{\mathrm{ATL}_{t-1}, \mathrm{ATL}_{t-10}\}$ and $\mathcal{P}_{\mathrm{Nino3}_{t-1}} = \{\mathrm{Nino3}_{t-2}, \mathrm{Nino3}_{t-4}, \mathrm{Nino3}_{t-6}\}$.

On the other hand, the peak at lag 1 for the cross correlation WEUR $\to$ EEUR in Fig. 6.2(b) is not much reduced and the value $\rho^{\mathrm{MIT}} = 0.1 \pm 0.07$ is even slightly larger than the link Nino3 $\to$ ATL with $\rho^{\mathrm{MIT}} = 0.09 \pm 0.06$, albeit this difference is negligible considering the large confidence bounds. The European time series have almost no autocorrelation which could alter the position and value of the peak. Additionally, we performed univariate regressions at the lag of the cross correlation's maximum and multivariate regressions using $\mathbf{U}^{\mathrm{MIT}}$ with the parents inferred by the algorithm as regressors. The results are summarized in Tab. 6.1. Also here, we see that the coefficient of the multivariate regression of ATL on Nino at lag 1 month is much smaller than that of the univariate regression at lag 4 (0.06 compared to 0.27) while the coefficients are unchanged in the European example within the error bounds. As can be seen from the R^2 values, Nino3 and ATL are rather well explained by their regressors, while the variance in WEUR and EEUR comes almost entirely from the innovation's variance, i.e., the source entropy. In Fig. 6.2(a) and (b), the black solid lines are the numerically evaluated cross correlations for a multivariate autoregressive process Eq. (2.14) with the same coefficients and innovations of Gaussian white noise with the same covariance matrix as the original residuals. The fitted lines well agree with the estimated correlations. This demonstrates, that a large part of the covariance structure can be explained by a Gaussian model based on the time series graph.

While the MIT values of the Nino3 $\to$ ATL and WEUR $\to$ EEUR links are equal within confidence bounds, the ITY value of Nino3 $\to$ ATL is significantly larger than the corresponding MIT value (0.22 ± 0.06 compared to 0.09 ± 0.06). As discussed in the analytical comparison in Section 5.2.1 and as expected from the inequality 5.2, the reason is that ITY becomes larger for strong autocorrelations within X, here within Nino3. A further difference is that some values that are significant for ITY became non-significant for MIT, for example, the auto-dependency link within Nino3 at lag 5. Again, the reason being, that the sample distribution of the partial correlation ITY is inflated for strong autocorrelations, see Section 4.3.1. In Sect. 6.5, we use a two-fold significance test as discussed in Sect. 4.4.4.

These differences between the measure ITY used in the PC algorithm and MIT are

further explored by the climatic examples in Fig. 6.2(c) and (d). Figure 6.2(c) shows,

Table 6.1.: Results of univariate and multivariate regression analyses (after subtracting the mean) and the covariance matrix of the residuals. The parents of every dependent variable in the time series graph are chosen as regressors, which can be read off the columns in Fig. 6.2(a) and (b). The coefficients of links relevant for the discussion are marked in bold.

Dep. Var.	Coeff. (lag τ)	Estim.	Std. Error	p-Value
Univariate regression				
ATL	**Nino3 (4)**	**0.27**	0.02	$< 10^{-5}$
Multivariate regression				
Nino3	Nino3 (1)	1.10	0.02	$< 10^{-5}$
	Nino3 (3)	-0.12	0.04	$< 10^{-3}$
	Nino3 (5)	-0.08	0.02	$< 10^{-3}$
				$R^2 = 0.91$
ATL	ATL (1)	0.83	0.02	$< 10^{-5}$
	ATL (10)	0.09	0.02	$< 10^{-5}$
	Nino3 (1)	**0.06**	0.01	$< 10^{-5}$
				$R^2 = 0.84$
	Residuals' covariance matrix			
		Nino3	ATL	
	Nino3	0.05	0.00	
	ATL		0.03	
Univariate regression				
EEUR	**WEUR (1)**	**0.18**	0.05	≈ 0.001
Multivariate regression				
WEUR	WEUR (1)	0.18	0.04	$< 10^{-5}$
	WEUR (3)	0.08	0.04	≈ 0.03
	WEUR (11)	0.08	0.04	≈ 0.02
				$R^2 = 0.05$
EEUR	EEUR (1)	0.24	0.03	$< 10^{-5}$
	WEUR (1)	**0.15**	0.05	≈ 0.004
	WEUR (12)	0.12	0.05	≈ 0.02
	WEUR (13)	0.14	0.05	≈ 0.01
				$R^2 = 0.10$
	Residuals' covariance matrix			
		WEUR	EEUR	
	WEUR	2.11	-0.02	
	EEUR		4.60	

that the precipitation rate over South America (SSA) has very low autocorrelations corresponding to a small coefficient b in our model example Eq. (5.1). For such low values we would not expect a shift of the peak of the cross correlation function which is also not the case here. But the exclusion of autocorrelation in Nino3 significantly reduces the value of MIT (0.19 ± 0.07) compared to ITY (0.29 ± 0.10) and the correlation (0.29 ± 0.08) of the link Nino3 $\to$ SSA lag 1. In Figure 6.2(d) we observe both cases in a feedback. The weakly autocorrelated WPAC drives (and is driven by) the highly autocorrelated EPAC at lag 1. The link WPAC $\to$ EPAC could have actually easily been overseen in a cross correlation analysis because it is not at the peak of the lag function. While here the values of ITY and MIT are almost equal (0.13 ± 0.06 for MIT versus 0.16 ± 0.06 for ITY), for EPAC $\to$ WPAC the value of MIT is much smaller than that of ITY (0.09 ± 0.05 versus 0.26 ± 0.06). Note that ITY does *not* entirely exclude autocorrelation in Y as shown in Tab. 5.1. In Appendix B.1 in Fig. B.1(a), we also show a 30-year sliding window analysis of this pair and note that the EPAC $\to$ WPAC link is observed more or less for the entire 1948–2012 period, while the WPAC $\to$ EPAC interaction becomes significant only from 1970 on.

Summarizing, we find a Nino3 $\to$ ATL link with a delay of one month rather than the broad peak around 4 months in the cross correlation, while for the other examples without strong autocorrelations in both variables the lag is – as expected – not shifted, but only the value differs apart from the weakly autocorrelated European time series. In Appendix B.2, we consistently find the same delay for the interaction between Nino3 and a region further in the north of the tropical Atlantic. We have tested the robustness of these examples by running the algorithm at different significance levels. As expected, the previously detected and more links occur for $\alpha < 95\%$-levels, at 97% the links Nino3 $\to$ ATL and WPAC $\to$ EPAC vanish, at 99% also EPAC $\to$ WPAC vanishes and at 99.9% also WEUR $\to$ EEUR becomes non-significant, while the strong Nino3 $\to$ SSA is still significant.

6.3.2. Climatological discussion

We now discuss the results of the previous section from a climatological perspective[10]. All results are shown in Fig. 6.3 along with the regions used in the analysis.

For the European example, we have found almost no difference between the lagged cross correlation and MIT, as expected due to weak autodependencies, i.e., low persistency, in the single time series. We have uncovered maxima for the European example cross correlation function at one month and 12–13 month time lag. The one month time lag well corresponds to quasi-stationary atmospheric planetary Rossby waves, which mediate this macro-turbulent synoptic scale heat exchange between Western Europe and Eastern Europe on the considered (monthly) time scale, with a pronounced seasonality inherent in these waves (Palmén and Newton, 1969).

[10]The interpretations in this section coming from Runge et al. (2014) were substantially contributed by co-author V. Petoukhov.

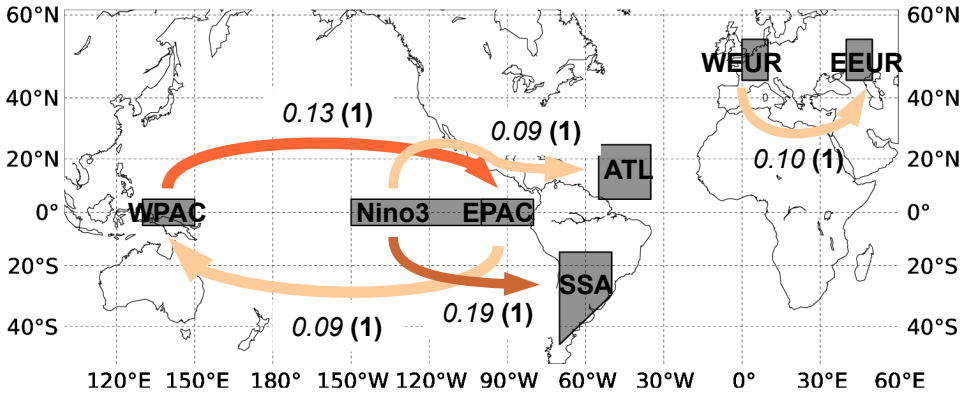


Figure 6.3.: Overview of important links determined in the analyses of ENSO teleconnections and in Europe. The gray boxes show the regions. The arrows indicate the direction with the shading roughly corresponding to the 'MIT strength'. The label gives the MIT value and time lag in months in brackets. Note that the 5% and 95% confidence bounds of these MIT values are typically ± 0.06.

Also for the influence of Nino3 on precipitation anomalies in southern South America, the peak of the cross correlation is at the same lag as the 'causal' link inferred in the algorithm. But here, the MIT value is smaller than the maximum of the cross correlation. This could be understood as an effect of the strong inertia in the tropical Pacific due to its large specific heat capacity. This implies that a large part of the co-variation between Nino3 and SSA is driven by a persistent momentum contribution from the past months in Nino3 due to the large oceanic heat capacity. MIT attempts to exclude these internal dynamics by "conditioning out" information in the past of both processes, resulting in a smaller value than the cross correlation. Still, the MIT value is the strongest coupling mechanism among the four studied bivariate examples.

For the Pacific – Atlantic teleconnection we have found that a model with a link Nino3 $\rightarrow$ ATL at lag one well explains the observed cross correlation function, which peaks at lag 4. A lag of about 3–6 months is also reported in many other studies, e.g., (Enfie and Mayer, 1997; Giannini et al., 2000; Wang et al., 2006; Chang et al., 2006). These studies also report higher peak values than we measure for the MIT. How can this difference be explained? What lags do the two approaches measure? In Sect. 5.5.3, we have given the simple picture of a particle Y fluctuating in a shallow potential well that models the internal dynamics. This particle is subject to random forces and to the external system X that acts on Y with a certain coupling delay τ via a coupling mechanism. In the Pacific – Atlantic teleconnection, this coupling mechanism corresponds to the "heat signal" being advected from Nino3 to the Atlantic region's atmospheric column by the Pacific – Atlantic Walker circulation. The characteristic horizontal velocity of this process is about $1-2\ \mathrm{m\,s^{-1}}$ (Wang, 2002) which well explains the delay of 1 month estimated in the time series graph for this

distance. After the signal arrives, the strong internal dynamics of the Atlantic in the oceanic mixed layer underlying the surface air counteract these "perturbations" with a characteristic time scale of about 3–6 months. This is the time delay quantified by the cross correlation, which measures the aggregate effect of the coupling mechanism plus the internal dynamics. Apart from the coupling delay, we have found that the MIT values of the Pacific – Atlantic teleconnection and the European teleconnection are the same suggesting that the coupling mechanisms via the Pacific – Atlantic Walker circulation and the synoptic macro-turbulence and planetary Rossby waves in Europe are actually of the same strength while the physics is very different. Note also that the Atlantic and Pacific regions are much further apart than the European regions studied.

6.4. Walker circulation

6.4.1. Statistical analysis

In the previous example, we have only considered the bivariate case of interactions between surface temperature anomlies in the East Pacific (EPAC) and surface pressure anomalies in the West Pacific (WPAC). Now we take into account a third time series of the average surface air temperature over a region in the Central Pacific (CPAC) shown on the map in Fig. 6.6 (150°–120°W, 5°S–5°N) and construct the time series graph for this three-variable process. Further, we compare the two-step approach run with partial correlations to conditional mutual information. This section is based on results published in Runge et al. (2014); Balasis et al. (2013).

First using partial correlation, to test whether the feedback between EPAC and WPAC was mediated via the surface of the central equatorial Pacific, we study the three-variable process (EPAC, CPAC, WPAC). Fig. 6.4 shows the analysis using the same significance level $\alpha = 0.95$ as before. The parents inferred are $\mathcal{P}_{\mathrm{W}_t} = \{\mathrm{W}_{t-1}, \mathrm{W}_{t-2}, \mathrm{W}_{t-10}, \mathrm{W}_{t-15}, \mathrm{C}_{t-1}, \mathrm{E}_{t-1}\}$, $\mathcal{P}_{\mathrm{C}_t} = \{\mathrm{C}_{t-1}, \mathrm{C}_{t-3}, \mathrm{E}_{t-1}, \mathrm{E}_{t-7}\}$ and $\mathcal{P}_{\mathrm{E}_t} = \{\mathrm{E}_{t-1}, \mathrm{E}_{t-2}, \mathrm{E}_{t-5}, \mathrm{W}_{t-1}\}$, where we abbreviated the variables by their first letter. Further, we found the contemporaneous links $\mathrm{E}_t - \mathrm{C}_t$ and $\mathrm{C}_t - \mathrm{W}_t$. Note that – as mentioned before – since the parents and neighbors are inferred with ITY, some of the corresponding links can have non-significant MIT values. The lagged cross correlation between EPAC and CPAC is broadly peaked around lag zero with a peak value of 0.75 ± 0.04 at lag 1. The MIT values are 0.32 ± 0.05 for the contemporaneous link and 0.15 ± 0.07 for the link EPAC $\to$ CPAC. It seems that the strong contemporaneous link prevents the peak from being shifted towards larger lags as would be expected for such strong autocorrelations. Note that the two links at lags zero and one are an example of a sidepath discussed in Section 5.2.3 and the MIT value at lag one, therefore, cannot be unambiguously related to this link. Further, CPAC drives WPAC with a lag 1. Very interestingly, the link EPAC $\to$ WPAC, that was robust before, vanishes. This result holds even for a low significance level of 95%. This link was obviously mediated via the surface of the equatorial Central Pacific. On the other hand, the link back WPAC $\to$ EPAC does *not* vanish (only

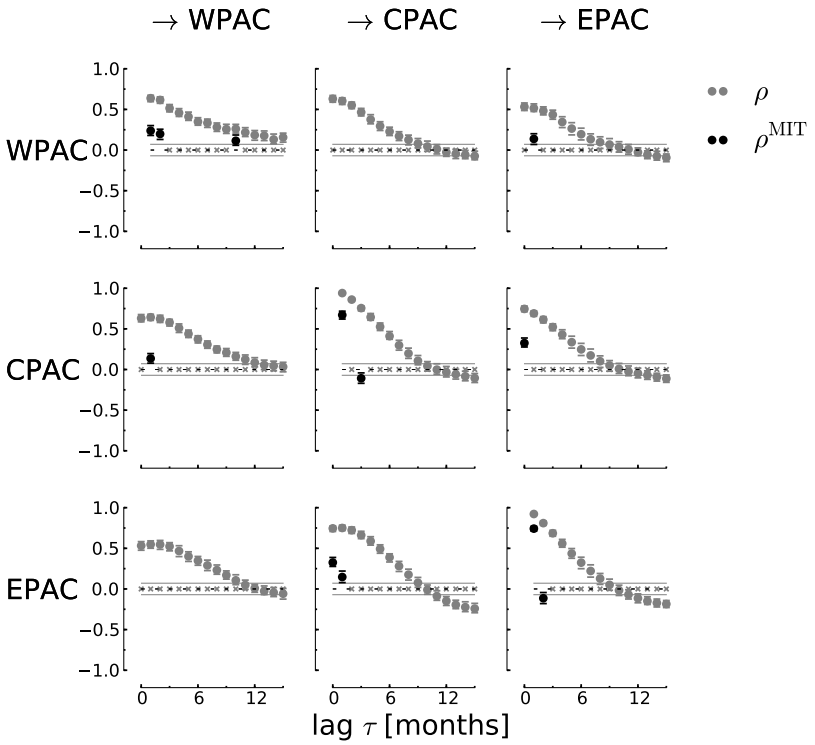


Figure 6.4.: Cross correlation (gray) and significant MIT values (black) for all pairs of variables (WPAC, CPAC, EPAC). For example, here MIT for the link CPAC → WPAC is the partial correlation $\rho(\mathsf{C}_{t-1}; \mathsf{W}_t | \mathcal{P}_{\mathsf{W}_t} \backslash \{\mathsf{C}_{t-1}\}, \mathcal{P}_{\mathsf{C}_{t-1}})$ with parents as given in the main text. The most important finding is the vanishing link EPAC → WPAC, which shows that the influence of the East on the West Pacific is mediated via the surface of the equatorial Central Pacific. On the other hand the link WPAC → EPAC stays, implying that this influence was not mediated via this region.

at higher significance levels) and the value is almost the same as in the bivariate example (0.14 ± 0.06). This shows, that the link back takes a different path, not via the surface of the equatorial Central Pacific. Also these results are recovered in a sliding window analysis as shown in Appendix B.1 in Fig. B.1(b). Interestingly, the MIT value of the link WPAC → EPAC along a distance of about 14,500 km is of the same strength as the CPAC → WPAC link with a distance of about 9,500 km.

Figure 6.5 shows the analysis using (conditional) mutual information for the same algorithm parameters and significance level, here obtained by a shuffle test as described in Section 4.3.3. We rescaled CMI to the partial correlation scale via Eq. (3.36) to make it better comparable. The MI lag functions are very similar to the cross correlation functions showing broad peaks. Also here, we find that the link EPAC → WPAC vanishes upon conditioning on CPAC. In contrast to the linear analysis, we find some more feedbacks, e.g., WPAC → CPAC and CPAC → EPAC, which

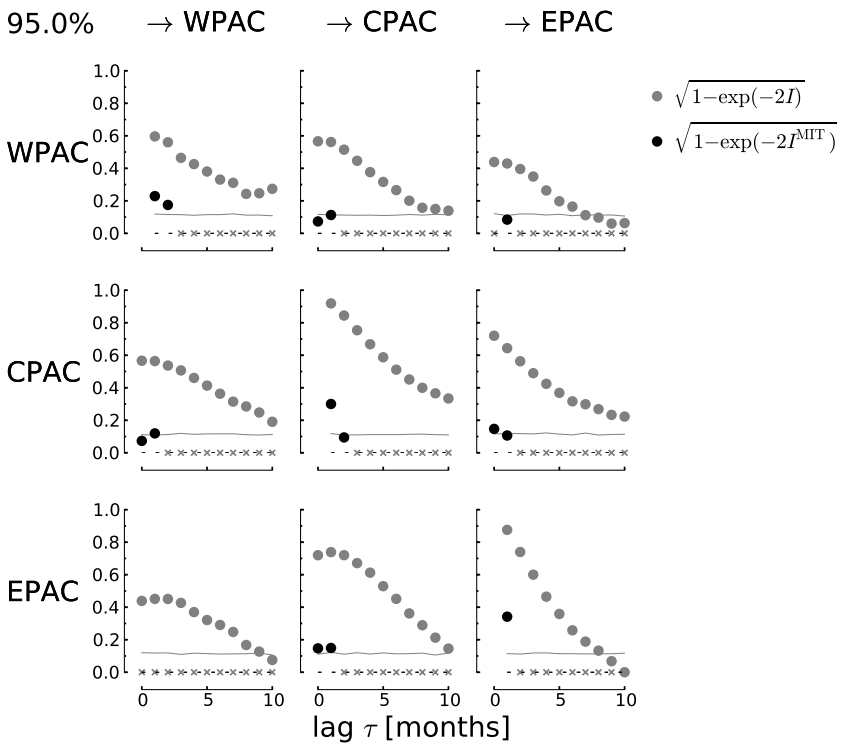


Figure 6.5.: As in Fig. 6.4, but for conditional mutual information. Note that the grey line denotes the significance level only for MI. The fact that some MIT values are below this line is a consequence of the bias for the higher dimensional CMI, but does not imply that the values are non-significant. As in the linear analysis, we find that the link EPAC → WPAC vanishes upon conditioning on CPAC.

actually can be climatologically understood as we will discuss in the next section. Summarizing, the trivariate example further demonstrates the power to detect indirect links not only in autodependencies (leading to shifted peaks) but also between multiple processes.

6.4.2. Climatological discussion

The basic mechanism of the Walker circulation (Walker, 1923; Walker, 1924; Bjerknes, 1969; Rowntree, 1972; Webster, 1981; Wang, 2002; Hosking et al., 2012) suggests that this circulation is primarily driven by heating on the western flanks of the equatorial oceans[11]. Figure 6.6, which is drawn on the basis of our results depicted in Fig. 6.4, illustrates well this feature for the Pacific branch of the above circulation. In normal and La Niña conditions, i.e., cold phases of ENSO, the latter is driven by strong sensible heating and latent heat release associated with penetrating moist convection

[11] The interpretations in this paragraph were substantially contributed by the co-author V. Petoukhov.

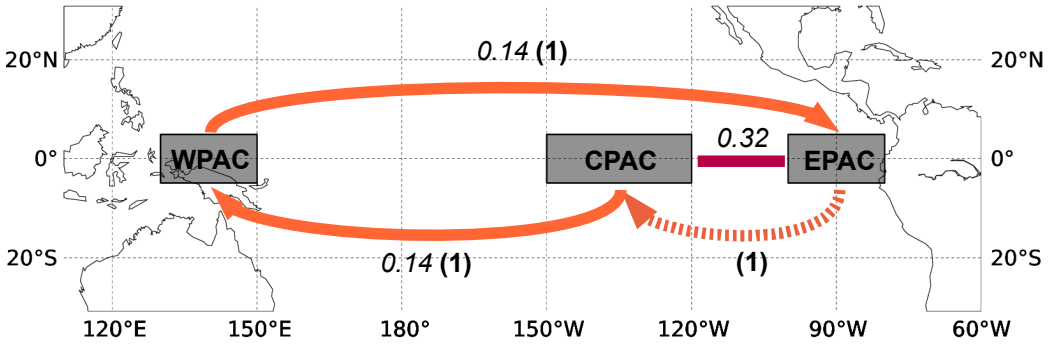


Figure 6.6.: Overview of important links determined in the linear partial correlation analysis in Fig. 6.4. The gray boxes show the three regions analyzed to study the Walker circulation. The arrows indicate the direction with the shading roughly corresponding to the 'MIT strength'. The undirected line EPAC − CPAC denotes the contemporaneous link and the link EPAC → CPAC is drawn dashed because this link has a sidepath via this contemporaneous link and the MIT can, therefore, not unambiguously be attributed to this link as discussed in Sect. 5.2.3. The labels give the MIT value and time lag in months in brackets. Note that the 5% and 95% confidence bounds of these MIT values are typically ± 0.06.

in the Western Pacific under a pronounced supply of the lower troposphere moisture there. The lower part of this circulation promotes upwelling of waters in the eastern part of the Pacific ocean and downwelling of waters in the western part. As far as oceanic temperatures decrease with increased depth, any decrease (increase) of surface pressure in the western part of the Pacific ocean, which accompanies an increase (decrease) in sea and air surface temperatures and moist convection there, favors a decrease (increase) of sea and air surface temperatures in the eastern and central Pacific. Our results track well this feature of the Pacific atmospheric and oceanic circulation: we have obtained a positive partial correlation MIT value between surface pressure over WPAC and surface air temperatures over EPAC. The link WPAC → CPAC that describes the descending mechanism is more pronounced in the analysis with conditional mutual information while it was missing in the linear analysis, which might hint at a nonlinear character of this dependency. As a confirmatory example, in appendix B.3 we study the vertical interaction between surface and tropospheric temperatures using conditional mutual information and find a strong vertical coupling for the ascending arm of the Walker circulation in the West Pacific. A mutual information analysis, on the other hand, shows strong associations throughout the tropics.

The described above Walker circulation pattern over the Pacific ocean is different during El Niño, i.e., warm events of ENSO, where the region of atmospheric updrafts shifts towards the Central Pacific and also broadens out. A statistical analysis of

the Walker circulation, thus, actually demands to analyze the different "seasons" (El Niño, La Niña and normal conditions) separately using the non-stationary time series graph approach from Section 2.4.5. Here we used the whole time sample to test the hypothesis that the "average" influence is mediated via the Central Pacific. The non-stationary approach discussed in Sect. 2.4.5 will be applied in the next Sections 6.5 and 6.6.

Summarizing, physically MIT is well interpretable as a measure that solely depends on the strength of a coupling mechanism and "filters out" internal dynamics, i.e., inertia or persistence, and even possible effects of external processes (if taken into account in the conditions). The strength of internal dynamics can be quantified by the corresponding auto-MIT value. The cross correlation and mutual information, on the other hand, cannot separate these influences. Both approaches measure different aspects of an interaction, but we believe that the improved interpretability of MIT is better suited to assist in understanding the underlying physics.

6.5. Interactions in global sea-level pressure system

In this section, we investigate causal interactions in a much larger complex system and employ more aggregated measures of interaction. In measuring link- and path-based interactions, we determine sources and sinks of information transfer and find that ENSO's outstanding role is manifest not only in its strong local (in the network) impact, but even more so due to its indirect effects as we infer with path-based measures of information transfer. This implies that local perturbations in ENSO can be measured strongly throughout the causal network, even in non-adjacent processes. In the next four sections, we explain the dimension reduction method and discuss interactions from a statistical perspective. In Sect. 6.5.5 we interpret these findings in a climatological context and discuss selected mechanisms.

6.5.1. Varimax components and time series graph estimation

We use the globally gridded dataset of surface pressures from the NCEP/NCAR reanalysis (Kalnay et al., 1996) for the period 1948 – 2012 on a monthly and weekly time scale. At a resolution of 2.5° in latitude and longitude, the data set consists of $10,512$ time series. Testing for statistical associations among thousands of pairs given comparably short data length of about 700 months (or about $3,000$ weeks) poses a serious estimation problem. In the statistics literature the estimation of high-dimensional covariance matrices is addressed using estimators where the number of non-zero entries in the covariance matrix is penalized *a priori* (Meinshausen and Bühlmann, 2006; Friedman et al., 2008). Additionally, the individual grid points are not the quantity of interest because they do not represent distinct climatological processes. As mentioned in the introduction to Chapter 2, such large datasets can also be analyzed by first reducing the dimensionality. Here, we follow this approach and construct components using *varimax rotated principal components* (Kaiser, 1958)

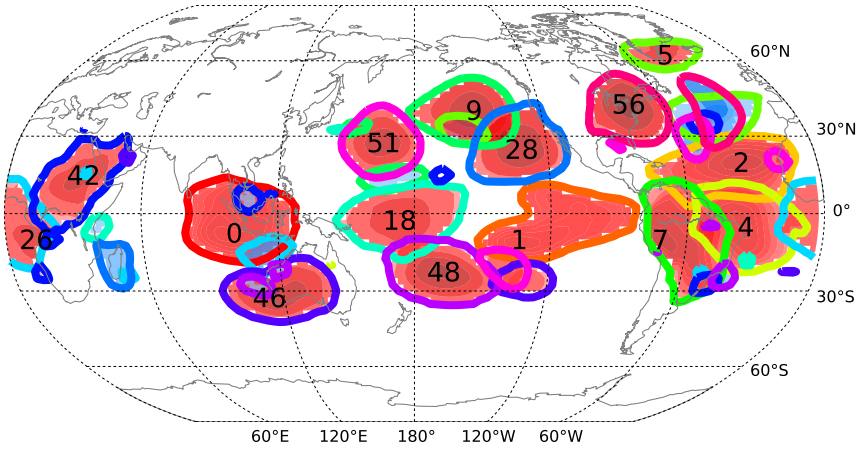


Figure 6.7.: Core (98% of the area-adjusted weight distribution) regions of weights (spatial loadings) for selected components. The color of the surrounding line identifies the parts belonging to one component. Some components have a dipole structure like No. 5, 9, and 56. Some can be associated with known climate indices like El Niño-Southern Oscillation (ENSO, No. 1), North Atlantic Oscillation (NAO, No. 5) or the Pacific/North American pattern (PNA, No. 9, more resembling the Pacific part of this pattern).

and a subsequent significance test arriving at 60 components that describe most of the variability in the data set (Hlinka et al., 2013)[12]. The rotation of principal components maximizes the sum of the variances of the squared principal component weights (loadings) and better represents regionally confined processes. Therefore, analyzing causal associations between these components allows to draw conclusions about interactions between climatic subprocesses rather than between grid points.

Here, we are interested in atmospheric interactions with time scales where dependencies typically decay within a month (Von Storch and Zwiers, 2002). To be able to infer directed interactions, we aggregate time series with a weekly resolution which provides a balance between too many statistical tests and not enough time resolution to conclude on causal directions. The data preprocessing steps to obtain the component weights are, however computed from the monthly gridded time series are as follows (Hlinka et al., 2013):

1. Anomalization in mean and variance, i.e., removing the mean annual cycle and dividing by the mean annual standard deviation

2. Linear detrending

[12]The varimax rotated components were provided by N. Jajcay and M. Vejmelka.

3. Rescaling according to latitude, i.e., cosine-transformed to account for the area a grid point represents (poles are excluded)
4. Estimation of covariance matrix (equal to the correlation matrix due to standardization)
5. Eigendecomposition of correlation matrix
6. Rotation using the varimax method (Kaiser, 1958)
7. Dimensionality reduction by comparing eigenvalues of original data (not components) to those from surrogate data preserving temporal structure, but destroying dependencies between the grid points, more precisely:
   a) Fit univariate auto-regressive model to each time series at each grid point separately using the Bayesian Information Criterion (Schwarz, 1978)
   b) Generate $10,000$ realizations of this model and obtain surrogate distribution of eigenvalues under this null hypothesis
   c) Find significant eigenvalues solving the multiple comparison problem using the False Discovery Rate (FDR) technique (Benjamini and Hochberg, 1995)

This procedure yields 60 significant components. In contrast to principal components, where the diagonal entries corresponding to the eigenvalues can be interpreted as the explained variability, for rotated principal components, the off-diagonal entries are not zero anymore and one cannot simply attribute an "explained variance" to each component. Still, we enumerate the components by the entry on the diagonal starting with the largest value (component No. 0). We have used monthly time series for the extraction of the components for computational reasons here and carrying out the decomposition directly on the daily time series might have provided a slightly different set of components, as the decomposition would also take into account high-frequency variability. Now the component weight matrix is multiplied with the weekly original gridded time series (that has been preprocessed by anomalization, standardization and cosine transform, steps 1–3 above) to obtain the weekly component time series. In Fig 6.7 we show the resulting component weights of selected components discussed in the following. Some components have a dipole structure like No. 5, 9 and 56. Some can be associated with known climate indices like El Niño-Southern Oscillation (ENSO, No. 1), North Atlantic Oscillation (NAO, No. 5) or the Pacific/North American pattern (PNA, No. 9). In the following plots we locate the component index at the location of maximum weight.

For these $N = 60$ processes we now estimate the time series graph up to a delay of $\tau_{\max} = 4$ weeks. Moreover, knowing that causal interactions are typically seasonal, we construct a non-stationary time series graph as defined in Sect. 2.4.5 using in the set of time indices $\mathcal{T}_Y$ only weeks that fall into the winter months November – April leading to about $1,700$ samples for each component time series. Here, we concentrate on the linear interactions, partially because the estimation using CMI would be

computationally too demanding, but mostly, because this allows to determine whether an interaction is 'anticorrelated' which is important for counteracting mechanisms as discussed in Sect. 5.2.4. Even though we have reduced the dimensionality considerably, the number of significance tests still poses a problem. For example, for a significance level as before of $\alpha = 95\%$ we would expect about $(1-\alpha)\cdot N^2\cdot\tau_{\max} = 720$ directed links simply due to chance. We therefore use a high (two-sided) significance level of $\alpha = 99.9\%$ which allows for about 15 spurious links, which will also be of a small correlation value. As discussed in Sect. 4.4, the α-level used in the PC algorithm is not a very reliable indicator due to the sequential testing problem. To overcome this problem, here we use the PC algorithm to estimate the parents and then test all possible links with MIT – taking advantage of the finding in Sect. 4.3.1 that MIT more faithfully fulfills the i.i.d. assumption –, and where we use the previously estimated parents as a conditioning set. In this way, we test each link only once and can more reliably assume to obtain 15 links by chance corresponding to a link density of 0.1% in the time series graph. Further parameters of the PC algorithm used here are $n_0 = 3$ (initial number of conditions), $n_{\max} = 8$ (maximum number of conditions checked), and $n_i = 5$ (number of tests per n), the latter to limit computational time.

We estimated lagged cross correlations and the partial correlations MIT and ITY, as well as the path-based measures ITP and MITP and linear interaction information IIP and its momentary version MII which can both also be negative. For IIP and MIT we use the absolute values of the partial correlations. Contrary to the previous applications where we discussed causal links in detail, here we discuss more aggregated network measures that quantify node statistics. In the following we refer by *network* to the process graph which aggregates the information in the time series graph by taking as a directed link between two processes only the lag with maximum absolute correlation. That is, the network nodes correspond to the $N = 60$ varimax components $\mathbf{X}$, not differentiating lags as for time series graphs, and edges and their lags and weights are aggregated from the time series graph by the rule

$$\mathbf{X}^i_{t-\tau}\to\mathbf{X}^j_t \text{ in time series graph for any } \tau>0 \implies i\to j \text{ in network edge set } E,$$
$$\tau_{ij} = \arg\max_\tau |\rho^{\mathrm{MIT}}_{i\to j}(\tau)|, \quad w_{ij} = \rho^{\mathrm{MIT}}_{i\to j}(\tau_{ij}). \tag{6.1}$$

All other w_{ij} are zero. We do not discuss contemporaneous links in the following since they do not allow for a causal interpretation. The resulting time series graph has a directed link density of 4% while that of the network is 17% with about 600 causal links. Remembering the general assumptions underlying our notion of causality, here saying "X causes Y" means *X Granger causes Y with respect to the pressure system components.*

6.5.2. Link strength

First, we look at local node measures. In Fig. 6.8(a), we show the number of significant causal, i.e., directed incoming and outgoing, links as the inner and outer node size, respectively. As mentioned before, we count only the strongest link between two

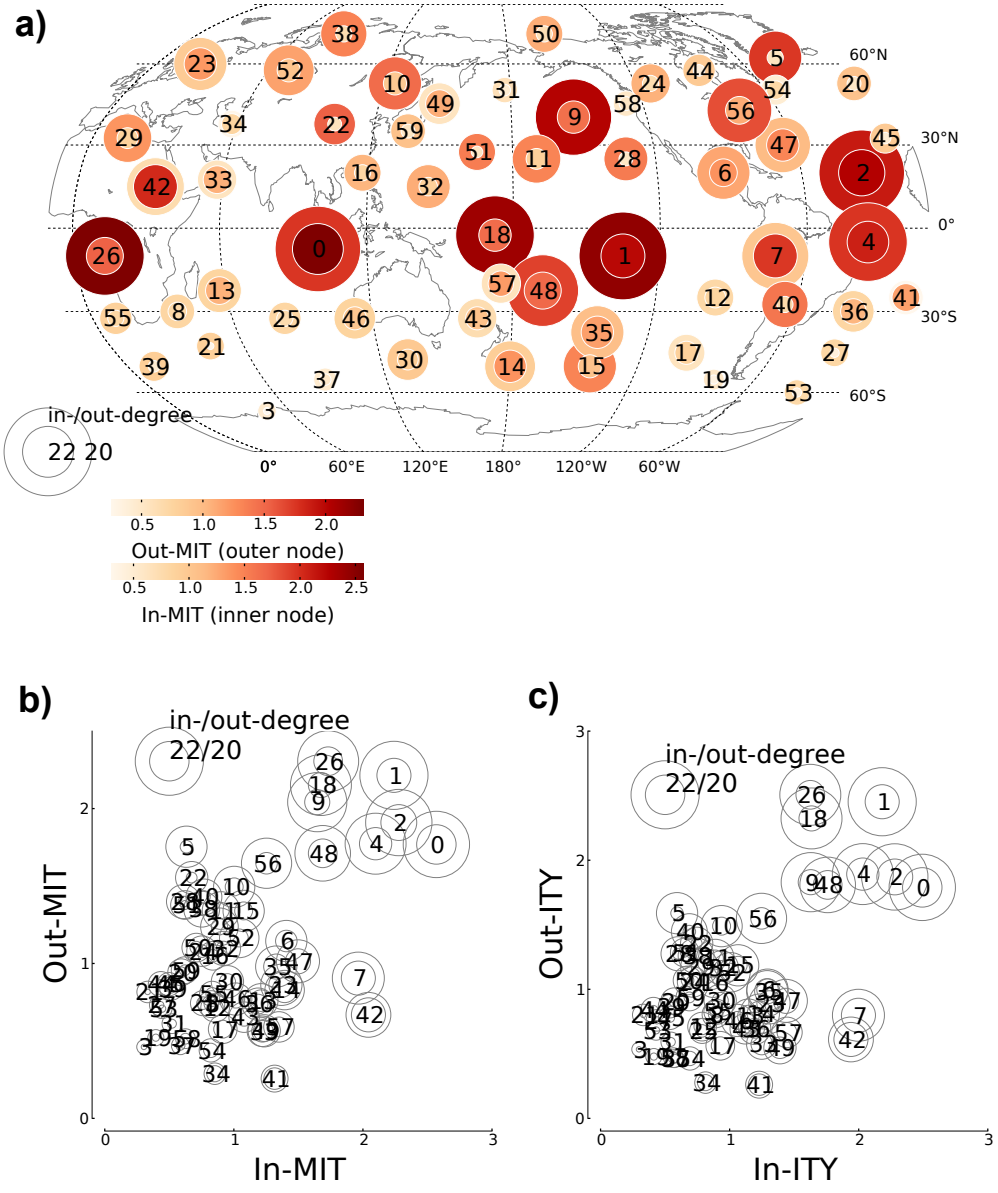


Figure 6.8.: Aggregated local node measures of causal network showing 'sources' and 'sinks' of information transfer defined in Eq. (6.2). In panels (a),(b), and (c) the inner and outer node sizes give the number of significant incoming and outgoing links, respectively, in the network. In (a) the corresponding node colors give the sum over the MITs of all these links, incoming (In-MIT) and outgoing (Out-MIT). In (b) these are plotted on the x- and y-axis and in (c) for In-ITY and Out-ITY.

nodes in the network (in the time series graph there could be multiple links for different lags). The maximum in- and out-degree then is 22 and 20, respectively, i.e.,

maximally a node directly influences about 20 other nodes in the network. Since this measure strongly depends on the chosen significance level just like the node degree in climate networks changes with the chosen threshold (Donges et al., 2009a), we show as a more robust node measure the sum – not the mean – over all incoming MITs (inner node color, In-MIT) and outgoing MITs (outer node color, Out-MIT), i.e.,

$$\begin{aligned} \text{Out}-\text{MIT} &= \sum_{j \in E} |\rho^{\text{MIT}}_{i \to j}(\tau_{ij})| \\ \text{In}-\text{MIT} &= \sum_{i \in E} |\rho^{\text{MIT}}_{i \to j}(\tau_{ij})|, \end{aligned} \tag{6.2}$$

where E is the edge set of significant links in the network and τ_{ij} the corresponding lag defined in Eq. (6.1). For Out-ITY and In-ITY, we sum about $|\rho^{\text{ITY}}_{i \to j}(\tau_{ij})|$. For better comparison, these measures are also plotted in a scatter plot in Fig. 6.8(b). In Fig. 6.8(c) we show the same scatter plot for ITY to demonstrate that both are qualitatively very similar. Note that pressure time series typically are much less autocorrelated than temperature time series (Sect. 6.3). In these scatter plots one can clearly distinguish "sources" and "sinks" of information transfer in the network. The group of nodes in the upper right corner consists mainly of processes in the tropical oceans (No. 0 in the Indian Ocean, No. 1, 9, 18, and 48 in the Pacific, and No. 2, 4, and 26 in the Atlantic). The dipole No. 56 is located above the US East Coast and also has strong In- and Out-MIT. Then there are nodes that are predominantly sinks (No. 7 in South America and No. 42 in East Africa) or sources (No. 5 in the North Atlantic).

The inferred time series graph contains information on the underlying physical mechanisms on which information can be transferred (assuming that the conditions formulated in Sect. 2.4.7 are correct). ENSO (No. 1), No. 26 south of West Africa, and No. 18 in the West Pacific have the strongest Out-MIT (also Out-ITY). This implies that perturbations entering the system through these processes are most strongly dispersed among the adjacent nodes which marks them as potentially affecting the stability of the system. But these perturbations are effective with a global influence only if they are measurable also in more distant processes of the network. In the next section, we study the global influence of processes along paths in the network that provide a notion of the efficiency of information transfer in this system.

6.5.3. Interactions along paths

Now, we investigate the information transfer not only between adjacent nodes, but also between nodes connected via causal paths as defined in Sect. 2.4.3. MITP and ITP are defined in Sect. 3.5.1. To limit computational cost, we estimated MITP and ITP for all pairs $\mathbf{X}^i_{t-\tau}$ and $\mathbf{X}^j_t$ that are linked by a *shortest path* of length 2 in the time series graph (not the network), i.e., $\mathbf{X}^i_{t-\tau} \to \mathbf{X}^k_{t-\tau_k} \to \mathbf{X}^j_t$, excluding pairs that are connected via a direct link. Further, we exclude feedback paths, i.e, we demand that i, k, j are mutually different, and check paths only up to a maximum total lag of

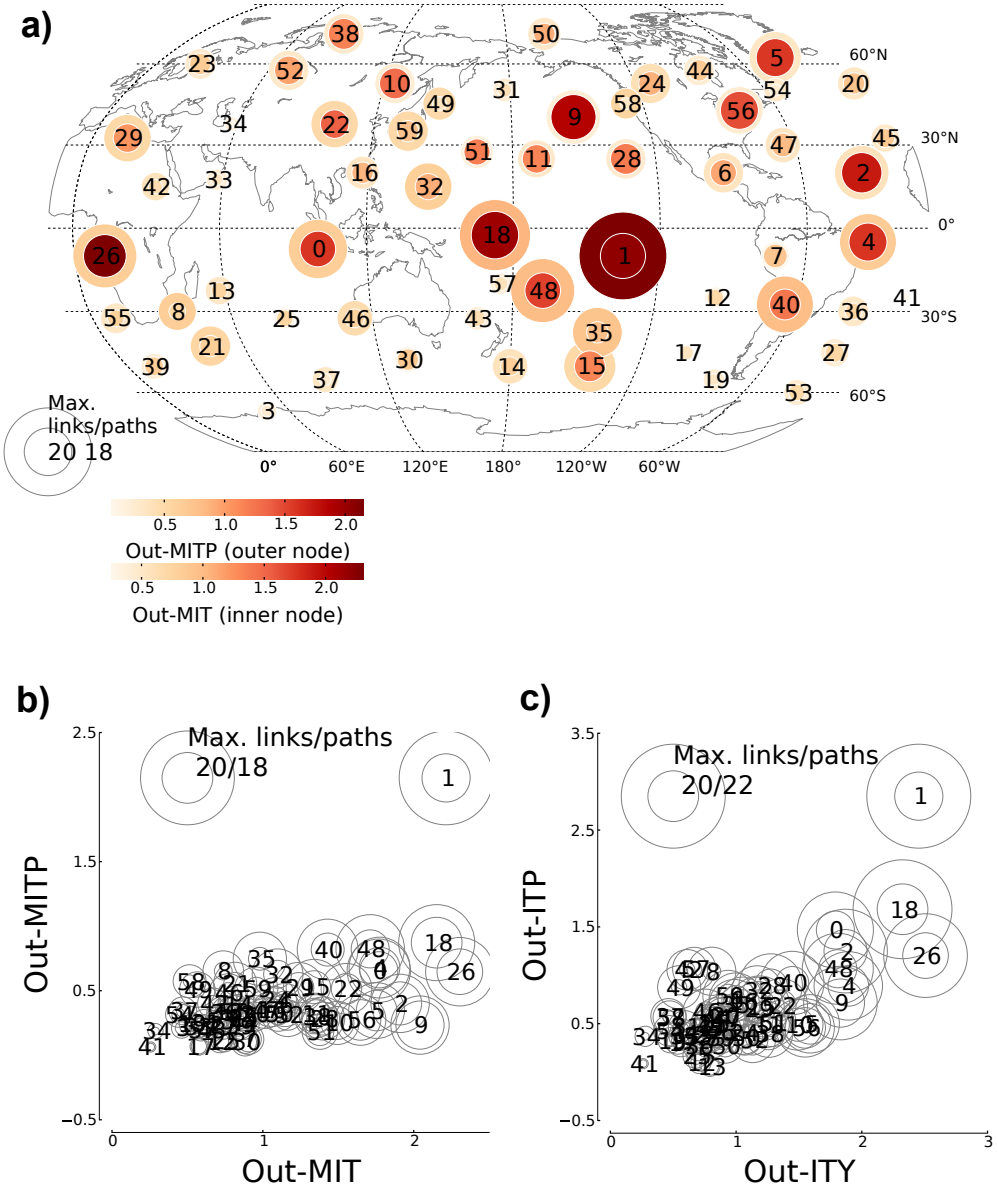


Figure 6.9.: Aggregated path-based node measures of causal network. In panels (a),(b), and (c) the inner node sizes give the number of significant outgoing links (like in Fig. 6.8 the outer node size) and the outer node size corresponds to the number of nodes a node is significantly associated with on causal paths (length ≥ 2) as measured by MITP. In (a) the corresponding node colors give the sum over the MITs and MITPs of all these links. In (b) these are plotted on the x- and y-axis and in (c) for ITP and ITY.

$\tau_{\max} = 4$. Note the difference between path-lengths and delays. Between each of these pairs, even though the shortest path is of length 2, there can be multiple causal paths of longer length. As defined in Sect. 3.5.1, then MITP is the information transfer between $\mathbf{X}^i_{t-\tau}$ and $\mathbf{X}^j_t$ along all of these paths conditioning on the parents of $\mathbf{X}^i_{t-\tau}$, $\mathbf{X}^j_t$ and of all nodes on all causal paths connecting $\mathbf{X}^i_{t-\tau}$ and $\mathbf{X}^j_t$, given by $\mathcal{C}_{X_{t-\tau}\to Y_t}$. In this way only the transfer of source entropy in $\mathbf{X}^i_{t-\tau}$ to $\mathbf{X}^j_t$ is measured. ITP, on the other hand, measures the transfer of any information entering $\mathbf{X}^i_{t-\tau}$, conditioning out only those parents of the end-node $\mathbf{X}^j_t$, that are not on any path. MITP, thus, measures the influence of 'perturbations' entering the whole system in $\mathbf{X}^i_{t-\tau}$.

In Fig. 6.9, we show the number of nodes in the network that a node has a significant MITP with – at any lag –, as well as the sum over all those significant MITPs, the aggregated Out-MITP, in comparison with the Out-MIT as discussed in the last section. For MITP and ITP we use a lower significance threshold of $\alpha = 0.95$. The node with the highest MITP degree, No. 1 (ENSO), indirectly influences 18 other processes in addition to the 19 processes directly influenced. As shown in Fig. 6.9(b), the summed up information transfer is by far larger than that of any other node. Several processes which strongly affect adjacent processes, have only very weak influence over paths, for example No. 9 (PNA), 2 (tropical Atlantic) and No. 5 (NAO). There is actually no clear correlation between Out-MITP and Out-MIT. If not only source entropy, but the transfer of all information entering $\mathbf{X}^i_{t-\tau}$ is taken into account (Fig. 6.9(c)), the correlation is higher. Still, also in this metric, ENSO is by far strongest. It implies that external or internal perturbations entering ENSO are distributed most strongly in the network. For example, the warm sea surface temperature anomalies developing during El Niño events in the East Pacific can be seen as an external perturbation of the sea-level pressure system which further propagates through the atmosphere along the causal information paths emanating from the ENSO component.

While we have investigated the path-based influence between two nodes in the network here, it would be interesting to know which nodes are important for transferring this information. This question will be analyzed in the next section.

6.5.4. Causal interaction betweenness

In this section, we measure the importance of an intermediate node on the causal paths discussed in the last section. Simply counting on how many causal paths a node is involved in, i.e., measuring the betweenness in the causal network, does not imply that the physical mechanism actually was mediated on this path as discussed in Sect. 5.5.4. For example, there are multiple paths between ENSO and NAO in the causal network, e.g., via No. 18 or No. 2, but still MITP and also ITP between ENSO and NAO are both zero. Actually even the simple cross correlation between the two is zero. This underlines the need for dynamic node measures that take into account the actual transfer of information. Here, we approach this question using interaction information defined in Sect. 3.5.2. For every pair with a significant ITP we measure by how much the ITP changes if a certain intermediate node is conditioned on. We

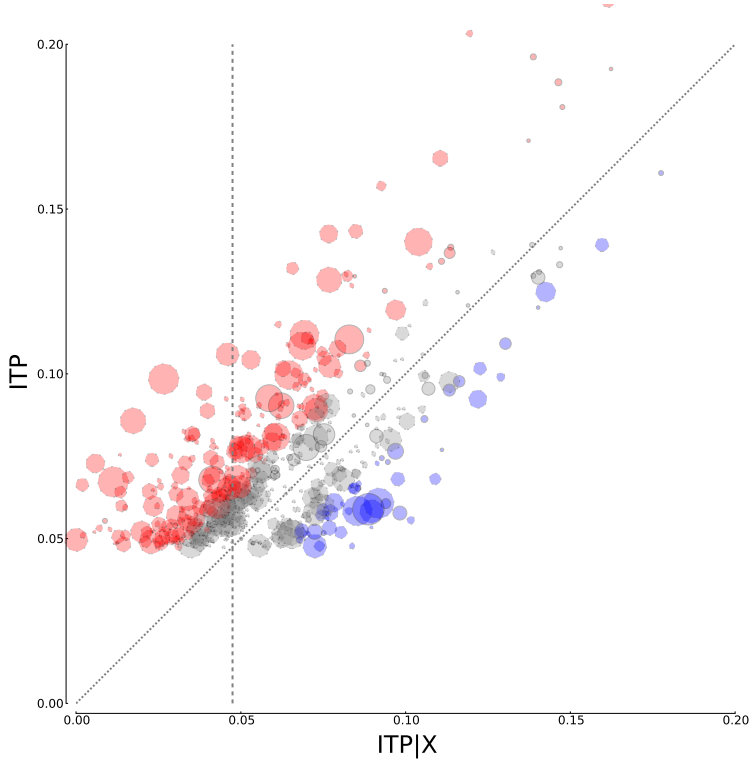


Figure 6.10.: Interaction information. Each circle corresponds to a pair $(\mathbf{X}^i_{t-\tau}, \mathbf{X}^j_t)$ with significant ITP as discussed in the previous section – here additionally taking into account directly connected pairs. The size of each circle corresponds to the number of intermediate nodes $\mathbf{X}^k_{\tau_k}$ with $k \neq i, j$ on any causal paths between the two. By a solid circle line we denote pairs that are additionally directly connected. On the y-axis we depict the ITP of each pair. For every intermediate process $\mathbf{X}^k_{\tau_k}$, we computed the ITP conditioned on $\mathbf{X}^k_{\tau_k}$. On the x-axis we show the value of this conditional ITP|X for the intermediate node for which the change was maximal. The interaction information $\mathcal{I}^{\text{IIP}}_{i \to j|k}$ defined in Sect. 3.57,1 then is the vertical (or horizontal) distance from the diagonal line which denotes triples with zero interaction. A significant positive change (enhancing interaction) of at least a standard deviation according to a Normal-Z test ($\alpha = 0.68$) is colored in red and a negative change (counteracting interaction) in blue, while smaller changes are in grey. There are almost no circles on the diagonal because we excluded pairs which are only directly connected without even a sidepath via intermediate nodes. The dashed vertical line denotes the significance level, circles left of this line mark non-significant ITP|X values where an intermediate process fully explains an interaction.

use ITP and not MITP because we are not interested where the information entered the system, but only how it is changed by the intermediate node.

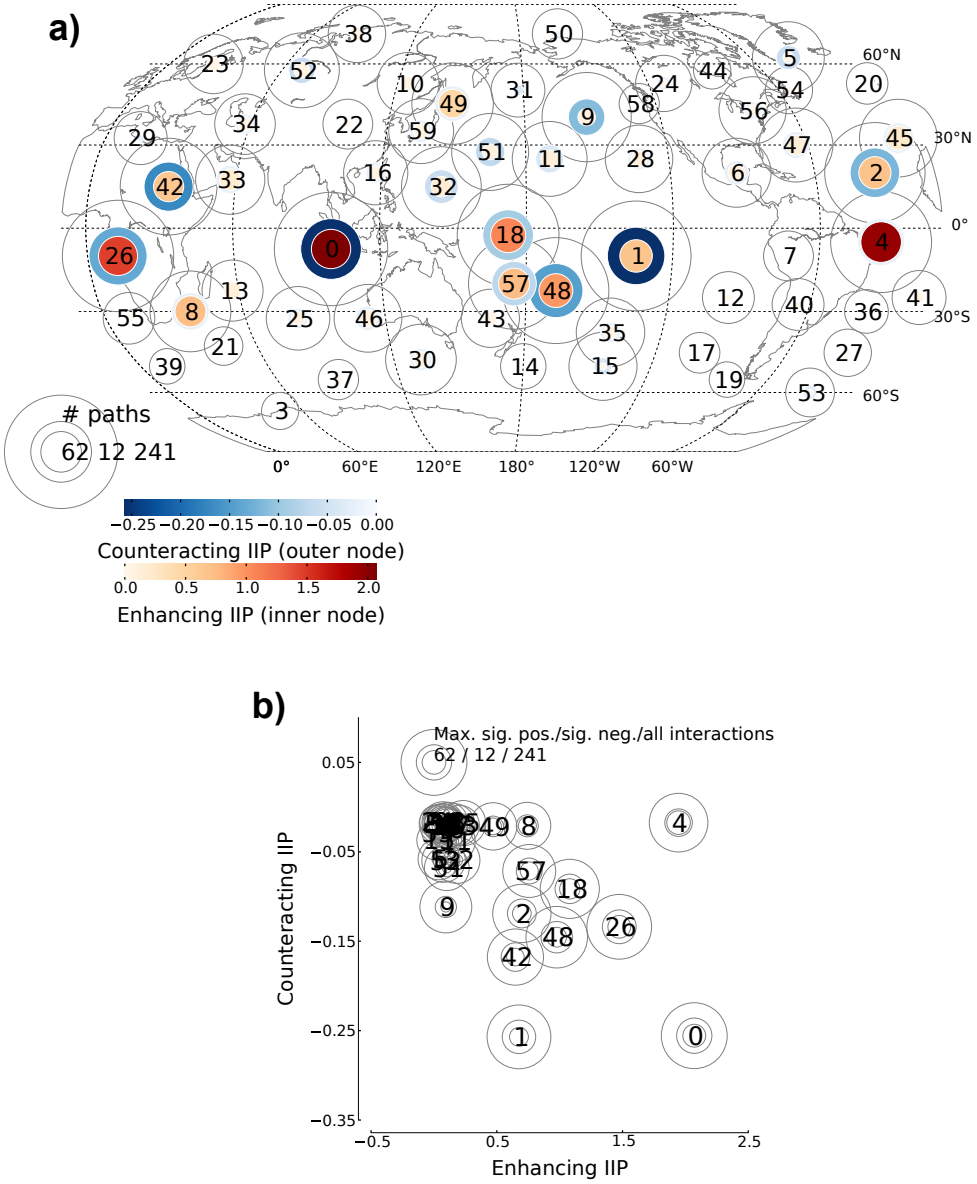


Figure 6.11.: Aggregated node interaction measures. (a), (b) The grey circle around each node depicts the (logarithmically scaled) number of causal paths through that process (summed over all its lags). The inner and outer node size depicts the number of significant interactions that this process is involved in. The inner (outer) node size shows the number of interactions where the process had a positive enhancing (negative counteracting) influence. (a) The inner and outer node color show the sum about all interaction informations. In the scatter plot in (b) the inner and outer node colors are plotted against the x- and y-axis.

Drawing on the analytical examples in Sect. 5.2.4, in Fig. 6.10 we show by how much the condition on intermediate processes increases (corresponding to a counteracting mechanism) or decreases (corresponding to an enhancing mechanism) the ITP between all pairs discussed in the last section – here additionally taking into account directly connected pairs. The red circles above the main diagonal denote enhancing interactions and the blue circles below counteracting interactions (the circle size corresponds to the number of intermediate nodes for this pair and we draw circles solid if the interaction also contains a direct link and dashed otherwise). For each pair we only plot the interaction for the intermediate node X for which ITP|X is maximally changed compared to ITP. The strongest interactions are mostly enhancing, that is, they explain part or even all of the ITP in which case the red circles lie on the left of the vertical line denoting non-significant ITP|X. We have not found interactions that enhance or counteract stronger ITPs which would be in the upper range of the scatter plot. For pairs with no direct link (dashed circles), i.e., with a zero MIT, this implies that only the combined effect of many intermediate nodes fully explains ITP, but no single node alone.

In Fig. 6.11, we show the interaction information aggregated for all processes, the causal interaction betweenness differentiating between counteracting and enhancing changes. Here, process No. 0 is most strongly enhancing as well as counteracting interactions. We discuss this process in the eastern Indian Ocean further in the next section. ENSO is equally counteracting, but enhances or explains interactions to a lesser extent. The betweenness centrality, discussed in Sect. 5.5.4, here is the number of causal paths passing through a node, here depicted by the size of the outermost grey circle in Fig. 6.11(a) and (b). But this measure does not take into account whether the node actually had any influence on an interaction. For example, even though No. 2 and 48 (and others) have the same number of causal paths passing through as ENSO, they are significantly influencing these paths to a much lesser extent. No. 9 (PNA) also here occurs as not very active. No. 4 in the tropical Atlantic, on the other hand is on par with No. 0 in enhancing mechanisms, but has almost no counteracting effect. In the next section, we investigate several interaction paths in more detail and discuss possible climatological mechanisms.

6.5.5. Climatological discussion of selected interactions

As already discussed in the climatological interpretations of the last sections, ENSO's impact on global climate is mainly due to *El Niño* and *La Niña* events. The former is characterized by an anomalous (compared to the mean seasonal climate) ocean surface warming that develops every 5 to 7 years in the central and eastern tropical Pacific. La Niña events, often following El Niño events, are of opposite sign, i.e., they refer to cold anomalies in the eastern tropical Pacific. Due to rising deep convection above the warmer ocean surface, these translate into negative pressure anomalies and vice versa for cold anomalies (Lau and Yang, 2003). The main climatological mechanism underlying ENSO's influence on the considered weekly atmospheric time scale is the "atmospheric bridge" (Alexander et al., 2004) by which the tropical atmosphere

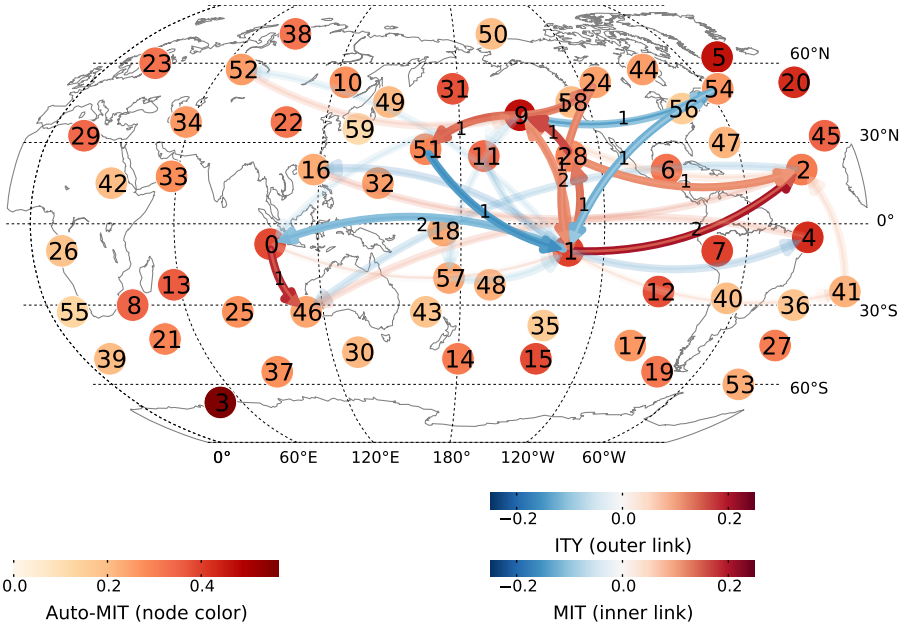


Figure 6.12.: Selected interaction paths. The node colors show the auto-MIT strength at lag 1. The links are divided into an inner and outer part. The inner (outer) link color denotes the MIT (ITY) strength of this link. Paths where the intermediate node has no significant influence on the interaction are transparent. For example, the positive influence between processes No. 1 and 2 is enhanced by No. 28 while the positive effect of No. 9 on 1 is counteracted via No. 51 and enhanced by No. 24 and 54.

responds to the early stages of ENSO events. Moist air from the eastern Pacific enhances precipitation across the tropical Pacific from the South American coast (compare to the Nino3 → SSA link found in Sect. 6.3) to the central Pacific. These anomalies are associated with diabatic heating which drives circulation changes also in the Indian and West Pacific Oceans.

For example, the indirect influence of ENSO on No. 46 in western Australia is strongly mediated by No. 0 (Fig. 6.12). The MITP of this indirect influence is (anticorrelated) -0.10 and the condition on No. 0 significantly reduces this value to -0.06. In Fig. 6.12 we show these links with the inner edge colored according to the MIT and the outer edge according to the ITY strength. The node color shows the autocorrelations which often lead to a large difference between MIT and ITY as seen in Sect. 6.3. There are several more causal paths between the two (transparent links), but only the one via No. 0 is significant here. In Fig. 6.11(b), we find that component No. 0 in the East Indian Ocean has the largest enhancing and counteracting influence in this complex system. It covers a region that defines the

eastern branch of the dipole mode index (DMI) (Saji et al., 1999). The DMI largely depends on ENSO, but has been claimed to play an independent role governing variability in Southeast Asia, Africa and Australia (Saji et al., 1999). Here, we find evidence for this hypothesis and confirm an analysis by Risbey and Pook (2009) who found that the DMI largely explains the correlation between ENSO and rainfall over Australia. They attribute this influence to DMI's impact on the subtropical jet stream which drives the development of synoptic systems over Australia.

The link of ENSO to the tropical Atlantic via the eastern branch of the Walker circulation was already discussed in Sec. 6.3. One such example shown in Fig. 6.12 is the direct link from ENSO to No. 2 in the northern tropical Atlantic at a lag of two weeks. This direct link is accompanied by several sidepaths (transparent links) of which only the one via No. 28 north of ENSO is significant. The MITP along all paths here is reduced by about 20% if No. 28 is conditioned out, implying that the direct link explains most of the interaction.

As an example of the influence of ENSO on the extratropical North Pacific, we study the influence on the PNA-component No. 9 at a lag of two weeks. Also here, No. 28 enhances this interaction. Here, we also observe a feedback from No. 9 to ENSO via paths through processes No. 0, 51, 11, 57, 24, and 54. The MITP over all these paths is not significant implying that perturbations entering the system trough No. 9 are not detectable in ENSO. Still, if all information entering No. 9 is taken into account a significant ITP of 0.10 is measured. Of all paths depicted in Fig. 6.12, only processes No. 51, 24 and 54 significantly change this ITP, i.e., they have a large IIP. Here, No. 24 and 54 act weakly enhancing (both reducing ITP to 0.08), while No. 51 counteracts and excluding its influence increases the ITP to 0.13. This mechanism consists of a strong positive influence of No. 9 on No. 51 and a further negative influence on ENSO leading to a net negative influence. The mechanisms via processes No. 24 and 54 have a net positive influence, where the mechanism via No. 54 has two negative links leading to a net positive effect. A possible mechanism for the northward influence of ENSO on No. 9 in the northern Pacific is by tropical convection leading to an overturning and subsidence within the Hadley circulation acting as a Rossby wave source (Trenberth et al., 1998). A similar feedback implying extratropical sea-level pressure anomalies influencing ENSO has also been studied in Vimont et al. (2002), albeit the 'seasonal footprint mechanism' proposed there acts on a longer time scale.

These exploratory findings can be seen as a preliminary step that needs to be substantiated by more detailed analyses taking into account more variables, seasonality and different kinds of ENSO events (e.g., via non-stationary time series graphs defined in Sect. 2.4.5). Still, our examples illustrate that very detailed insights on interactions in complex systems can be retrieved from such an analysis.

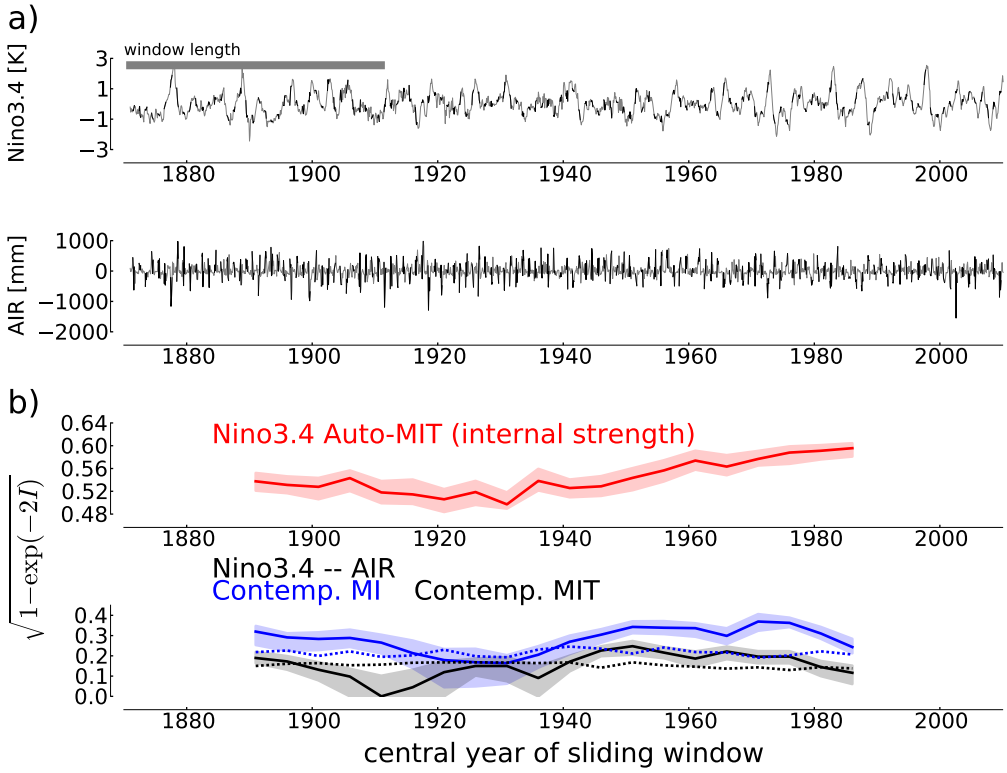


Figure 6.13.: Interactions between ENSO and the Indian Summer Monsoon using the Nino3.4 and All India Rainfall (AIR) indices. (a) Anomaly time series with the June – September monsoon season marked in black. (b) Time dependent values of the Nino3.4 auto-MIT (red) and the contemporaneous MIT (black) and MI (blue). The time index denotes the center of the sliding windows (marked by the grey bar in (a)). All (C)MIs were estimated with $k = 30$ and rescaled to the correlation scale using Eq. (3.36). The shaded interval denotes the 70% confidence region assessed from a bootstrap test (Sect. 4.3.4). The dotted lines denote the 95% shuffle test significance levels for MI (blue) and MIT (black). The level for MIT is lower due to a smaller variance for this higher dimensional CMI (Sect. 4.3.3).

6.6. Time-dependent interactions between El Niño-Southern Oscillation and the Indian Summer Monsoon

The Indian Summer Monsoon (ISM) is of major importance for agriculture in India. Determining its drivers can help understand its failure in some seasons and may lead to improved predictions. The interconnections between ENSO and the ISM have been extensively studied concluding on a persistent statistical relationship with warmer ENSO anomalies being associated with weaker monsoon rainfall anomalies in the monsoon season June – September (Pant and Kumar, 1997; Kumar et al.,

1999). Next to linear analyses, also studies with phase synchronization and nonlinear model-based Granger causality found mutual dependencies (Maraun and Kurths, 2005; Mokhov et al., 2011). The common mechanism explaining this relationship is through an eastward shift of the Pacific Walker circulation. During an El Niño event the ascending limb of this circulation is shifted from the West to the Central Pacific. This shift leads to an anomalous subsidence extending to the Indian Ocean, which suppresses convection and precipitation there. Here, we investigate the dependence between the monthly Nino3.4 and All India Rainfall (AIR) indices (Parthasarathy et al., 1994). Nino3.4 is constructed from averaging sea surface anomalies (Rayner et al., 2003) over the region (170°–120°W, 5°S–5°N), which extends more to the central Pacific than the region Nino3 analyzed in Sect. 6.3. Kumar et al. (2006) have shown that the Central Pacific is more effectively leading to the above mentioned subsidence over India. In Fig. 6.13(a) we depict the time series.

In the introductory chapter, we have mentioned that studying the time-dependence of interactions can help determine possible tipping points (Lenton et al., 2008), i.e., structural changes of the dynamics, here of the causal interactions. To this end, we investigate the time dependent nonlinear interactions with MIT in a sliding window analysis of window length 40 years (480 months). Applying the PC algorithm restricted to the monsoon season June – September, we infer a time series graph for every sliding window. Across all windows only the contemporaneous link Nino3.4 – AIR and an autodependency link in Nino3.4 at lag 1 are robust. Even though a contemporaneous link does, in general, not allow for a causal interpretation from a statistical perspective, the physical mechanism described above and climate model simulation studies (Kumar et al., 1999) provide evidence to speak of a causal relation here that acts on a smaller time scale than the monthly time-resolution used. In Fig. 6.13(b), we show the time-dependent strength of these two MITs together with the contemporaneous mutual information (MI). During 1900–1940, both MI and MIT are rather weak with the MIT value even below the significance threshold consistent with other studies, e.g., using a phase coherence measure in Maraun and Kurths (2005). From the 1940s on, we observe a steadily increasing auto-MIT, which we have argued in Sect. 5.5 to measure the 'internal strength', here of ENSO. Co-occurring with this increase, MI shows an upward trend up to around 1970–80. In Sect. 5.2.1 we have seen that the correlation and hence also MI strongly depend on the auto-MIT of the driver, which explains this increase here. MIT, on the other hand, is more or less stationary with a decrease from 1965 on. From around 1975 on, both, MI and MIT, feature a strong decrease, even to non-significant values in the case of MIT. Note that the given time periods refer to the center of the sliding windows.

Kumar et al. (1999) attribute the weakening relationship found also with linear correlation measures in recent decades to a southeastward shift of the Walker circulation (additionally to the normal shift during ENSO events) on the one hand, and greenhouse-gas induced increased surface temperatures over Eurasia on the other hand. These, increasing the sea-land temperature contrast between the Indian Ocean and the Himalayan plateau, sustained the basic monsoon mechanism and can, thus, be argued to have prevented Monsoon failures in recent decades despite several strong

ENSO events (Kumar et al., 1999).

The unprecedented and abrupt (even more visible in shorter sliding windows, which do, however, not allow for a reliable estimate of MI anymore) change in the coupling mechanism between ENSO and ISM can be seen as as a tipping point and the fact that MIT indicates this change earlier than MI, independent of ENSO's internal strength, provides some evidence that MIT might be a better proxy of an order parameter possibly underlying this process. Note, however, that many different kinds of tipping point mechanisms exist, each requiring different early warning indicators (Scheffer et al., 2009; Thompson and Sieber, 2011a; Thompson and Sieber, 2012; Thompson and Sieber, 2011b; Kuehn, 2011) and it is unclear to which type the interaction between ENSO and ISM belongs to. An interesting further perspective here is to include other processes such as the Indian Ocean Dipole (Saji et al., 1999; Ashok et al., 2004) in an analysis to obtain a more refined picture as shown in the last section. This preliminary analysis demonstrates the potential use of our approach to determine critical transitions in complex coupling mechanisms.

6.7. Summary

In this chapter, we have extensively demonstrated the novel methods on geophysical datasets, using surface pressures, temperatures and precipitations from daily weather to monthly climate time scales.

In a first step (Sect. 6.3), we analyzed interactions between two processes for which autocorrelations can misguide a physical interpretation. For the influence of the tropical East Pacific on the northern tropical Atlantic, we detected a short lag of one month for this coupling mechanism consistent with the advection speed of the Pacific – Atlantic Walker circulation, while previous studies using the maximum of the cross correlation lag function found lags of 3–6 months. Also, we unveiled that the coupling mechanism is actually quite weak (even comparable to the coupling mechanism between Western and Eastern Europe) and that the large cross correlation value can be explained by strong autocorrelations present in both time series. As a further step (Sect. 6.4), we investigated three processes to validate our method on the mechanism of the Walker circulation. The purely statistical analysis confirms that the positive correlation of surface temperatures over the Eastern Pacific and surface pressure over Western Pacific is mediated via the Central Pacific while the lagged correlation back cannot be explained by variabilities in surface temperatures of the Central Pacific. The time lags of this circulation are weeks to one month between the Eastern and Central Pacific, another month for the impact of the Central Pacific on the Western Pacific and one month for the link back via the upper atmosphere. For the path CPAC $\rightarrow$ WPAC $\rightarrow$ EPAC, we find that the strength of these two mechanisms is very similar even though they act on very different distances. These examples demonstrate that the methods enable climate researchers to statistically

test specific hypotheses on interactions in the data. While the concept introduced here is purely statistical, it may serve as a first step to construct conceptual or more complex models of physical processes.

In the framework of exploratory data analysis, we have applied the novel measures on a global pressure dataset consisting of 60 components that represent distinct climatological subprocesses (Sect. 6.5). We used measures to quantify not only directly linked mechanisms, but also indirect ones via paths and the interaction between multiple processes. As argued in Sect. 5.5.4, our method provides a complementary approach to climate networks (Donges et al., 2009a) that takes into account the actual causal transfer of information, also indirectly via paths. We found that many subprocesses in the tropical oceans have strong influence on adjacent nodes in the causal network, while ENSO's influence can be strongly measured even on processes that are not directly causally linked with ENSO. Further, we demonstrated that interaction information can be used as an alternative to betweenness centrality that takes into account the dynamic causal transfers of information. In selected examples, we determine and climatologically discuss the mechanism by which ENSO influences West Australia, the tropical Atlantic and the North Pacific. Summarizing, the novel introduced metrics of path interactions allow to characterize very precisely how interactions between distant (in the network sense) nodes are mediated and aggregated node measures allow to determine the importance of single processes in a complex system as sources and sinks and mediators of information transfer.

In a sliding window analysis of the interaction between ENSO and the Indian Summer Monsoon (Sect. 6.6), we demonstrated the potential use of our method to determine critical transitions in the strength of causal mechanisms. Here, the decoupling between ENSO and the Indian Monsoon was detected earlier with MIT compared to MI. However, we have argued that finding good 'early-warning' measures sensitively depends on the type of critical transition (Scheffer et al., 2009; Kuehn, 2011). While we have given only a preliminary example here, the study of critical transitions constitutes an interesting and important area not only in climate research and conditional measures such as introduced in this thesis might turn out to be of particular use for this task.

This chapter demonstrated that the novel methods can help in understanding mechanisms in the climate system. Causal discovery methods in climate research have so far mainly been used in specialized applications such as decision-making tools (Ebert-Uphoff and Deng, 2012b), and the methods introduced here are among the first pioneering works in this area after their introduction in Ebert-Uphoff and Deng (2012a); Ebert-Uphoff and Deng (2012b). We have studied our approach in the linear as well as the nonlinear framework for some examples and have found consistent results with some deviations that might hint at nonlinear relationships (in the Walker circulation example Sect. 6.4). This similarity is often due to the considered spatial and temporal scales on which nonlinearities are largely averaged out leading to rather linear observations. However, on smaller spatial and temporal scales and for other climatic variables, such as precipitation, we expect stronger nonlinearities. Further, we found that the exclusion of autocorrelation effects is of particular importance in

climate time series. Autocorrelations are an ubiquitous feature of time series also in many other fields, for instance in economics and neuroscience, and our approach to overcome autocorrelation in the detection as well as the quantification of associations could be utilized also in these fields.

The detection of causal mechanisms can be further used to predict a system's future dynamics as shown in Chapter 7.

Chapter 7.

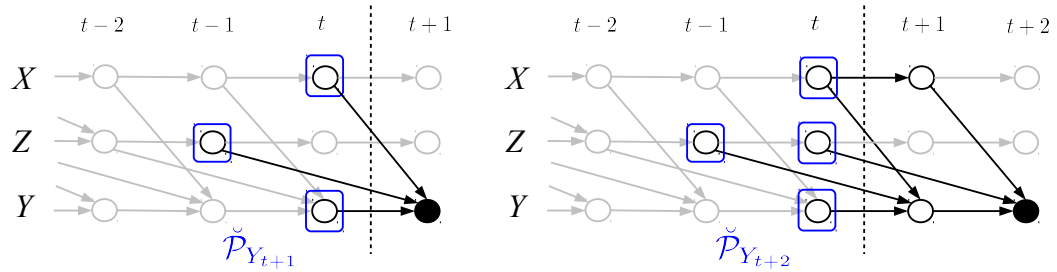
Time series prediction

7.1. Introduction – from causality to prediction

While the inference of causal interactions constitutes a goal in itself, we already mentioned earlier that it can be seen as a first step to build a model and predict a complex system. For example, El Niño-Southern Oscillation (ENSO) with its far-reaching climatic and economic impacts has been the focus of prediction research for many decades starting with the "Cane-Zebiak" model (Cane et al., 1986; Zebiak and Cane, 1987). Since then, forecasts have steadily improved on two branches: using climate models and statistical prediction with the forecast skill of the former outperforming statistical predictions only in recent years. Statistical predictions employed are mostly linear regressions (Latif et al., 1998; Barnston et al., 2012) using principal components of climatological fields from sea surface temperature and other variables. Also the recent approach to combine complex network theory with correlation networks as discussed in Sect. 5.5.4 has been used to predict ENSO (Tsonis and Swanson, 2008; Ludescher et al., 2013; Ludescher et al., 2014).

But since the late 1980s also model-free predictions have been developed using nearest neighbors in state space (Farmer and Sidorowich, 1987; Abarbanel et al., 1990; Giona et al., 1991; Alparslan et al., 1998; Ragwitz and Kantz, 2002) or neural networks (Eisenstein et al., 1995; Szpiro, 1997; Small and Tse, 2002). In the nearest-neighbor technique, states similar to the present state are searched for in the past of the time series and a future value Y_{t+h} at a prediction step h is forecasted by simply averaging the past values corresponding to the nearby past states or using local-linear models (Farmer and Sidorowich, 1987). The difference between these two is that the former will only produce values in the range that already occurred while the latter can also extrapolate. Nearest-neighbor techniques have also been used in weather forecasting (Yakowitz and Karlsson, 1987).

In a univariate setting, in these methods states are usually reconstructed from embedding the time series using Taken's theorem (Takens, 1981; Ragwitz and Kantz, 2002), but also here the curse of dimensionality has hampered the use of multivariate predictions. In this chapter, we investigate how the knowledge of the causal parents of a process can provide an optimal scheme for prediction. In Groth (2001); Pompe (2002) a similar optimization scheme has been investigated. Also here, the prediction using these optimal causal predictors can be performed using linear or model-free prediction. In the following sections, we prove that causal parents yield optimal



a) Optimal predictors for Y_{t+1} **b)** Optimal predictors for Y_{t+2}

Figure 7.1.: Optimal predictors in example time series graph. For the one time step ahead prediction in (a) the optimal predictors are simply the parents. In (b) for predicting Y_{t+2}, the processes Y_{t+1} and X_{t+1} already lie in the future and are not available anymore while Z_t is still observed. Part of the information in Y_{t+1} and X_{t+1} can still be recovered by measuring their parents X_t and $\{Y_t, Z_{t-1}\}$, respectively, which share information along the paths marked with black arrows. Generally, optimal predictors are found by determining the Markov set of nodes as discussed in the text.

predictors, discuss the problem of coping with *overfitting*, i.e., fitting the predictors too much to the sample rather than the process, and apply our scheme to the prediction of ENSO. We find that the prediction of ENSO can be considerably improved, especially if the optimizing scheme is applied in the linear framework.

7.2. Optimal prediction

7.2.1. Optimal predictors

From the information-theoretic perspective, the optimal predictors of a process $Y \in \mathbf{X}$ at a prediction step $t+h$ for $h > 0$ are those $\mathcal{S}_{t+1} \subset \mathbf{X}^-_{t+1}$ with $\mathbf{X}^-_{t+1} = (\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots)$, that maximize the multivariate mutual information $I(\mathcal{S}_{t+1}; Y_{t+h})$ (Groth, 2001; Pompe, 2002). For $h = 0$, i.e., "predicting" the present, it is easy to see that this set must contain the parents $\mathcal{P}_{Y_t}$ defined in Sect. 2.4.3, because they separate Y_t from the rest of the process $\mathbf{X}^-_t \backslash \mathcal{P}_{Y_t}$ in the time series graph (Markov property). To prove this, we denote by $\tilde{\mathcal{P}}_{Y_t} \subset \mathbf{X}^-_t$ a set that "misses" some parents, i.e., $\mathcal{P}_{Y_t} \backslash \tilde{\mathcal{P}}_{Y_t} \neq \emptyset$ and use the chain rule for a multivariate MI as follows:

$$I\left((\tilde{\mathcal{P}}_{Y_t}, \mathcal{P}_{Y_t}); Y_t\right) = I(\tilde{\mathcal{P}}_{Y_t}; Y_t) + \underbrace{I(\mathcal{P}_{Y_t} \setminus \tilde{\mathcal{P}}_{Y_t}; Y_t | \tilde{\mathcal{P}}_{Y_t})}_{>0 \quad \text{(parents)}} \tag{7.1}$$

$$= I(\mathcal{P}_{Y_t}; Y_t) + \underbrace{I(\tilde{\mathcal{P}}_{Y_t} \setminus \mathcal{P}_{Y_t}; Y_t | \mathcal{P}_{Y_t})}_{=0 \quad \text{(Markovity)}}, \tag{7.2}$$

from which it follows that $I(\tilde{\mathcal{P}_{Y_t}}; Y_t) < I(\mathcal{P}_{Y_t}; Y_t)$ proving the optimality of the parents $\mathcal{P}_Y$ as predictors of Y. This actually holds for any set $\mathcal{S}_t$ that contains the parents as a subset, i.e., $\mathcal{P}_{Y_t} \subset \mathcal{S}_t$, because

$$I(\mathcal{S}_t; Y_t) = I(\mathcal{P}_{Y_t}; Y_t) + \underbrace{I(\mathcal{S}_t \setminus \mathcal{P}_{Y_t}; Y_t | \mathcal{P}_{Y_t})}_{=0 \quad \text{(Markovity)}}, \tag{7.3}$$

and the parents are, therefore, the *minimal set of optimal* predictors. For $h > 0$, the minimal set of optimal predictors $\breve{\mathcal{P}}_{t+h} \subset \mathbf{X}^-_{t+1}$ of future values Y_{t+h} also needs to separate Y_{t+h} from $\mathbf{X}^-_{t+1} \setminus \breve{\mathcal{P}}_{t+h}$. These optimal predictors encode *Granger causality at a time horizon* $h > 0$ (Sims, 1980; Hsiao, 1982; Dufour and Renault, 1998) and are defined as

$$\breve{\mathcal{P}}_{t+h} \equiv \{X_{t-\tau} \in \mathbf{X}^-_{t+1}, \tau \geq 0 : X_{t-\tau} \not\perp\!\!\!\perp Y_{t+h} \mid \mathbf{X}^-_{t+1} \setminus \{X_{t-\tau}\}\}, \tag{7.4}$$

where $\mathbf{X}^-_{t+1} = (\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots)$. We mark this set with a breve to distinguish it from the notion of parents $\mathcal{P}_{t+h}$ as defined in Eq. (2.10), which would also include future processes not available as predictors. In Fig. 7.1 we give an example. The optimal predictors can also be described as the first ancestors of Y_{t+h} before the point $t+1$ in time. Typically, but not necessarily, the number of predictors will grow with the ahead step h. Because the PC algorithm introduced in Sect. 2.4.6 can consistently infer the separating (Markov) set of processes, we can also use it to estimate $\breve{\mathcal{P}}_{t+h}$ with the restriction that we search for parents only within $\mathbf{X}^-_{t+1}$. The only uncertainty left then comes from the source entropy of Y_{t+h} plus the entropy from the unobserved ancestors of Y_{t+h} between $t+1$ and $t+h-1$.

7.2.2. Prediction scheme

In general, our prediction scheme consists of three steps performed separately for each prediction step ahead $h = 1, \ldots, h_{\max}$:

1. Estimate predictors $\breve{\mathcal{P}}_{t+h}$ from the observed time series with the PC algorithm.
2. Rank predictors using forward selection.
3. Forecast the unobserved future value Y_{t+h} using nearest-neighbor prediction.

In the following, we describe the details of these steps. In a linear framework, the last step can be substituted by an autoregressive prediction and the first two steps are performed with partial correlation instead of conditional mutual information.

With the PC algorithm described in Sect. 4.4, we obtain a set of predictors for a given significance level or fixed threshold. Albeit this set theoretically allows for an optimal prediction, the estimated predictors can be a result of *overfitting*, which means that they do not actually describe a causal relationship, but merely fit the noise in the data. This problem can be addressed using heuristics to penalize the addition of predictors or by *cross-validation*. Here we employ a *forward selection*

procedure to rank the predictors and use a heuristic criterion for the optimal number of predictors. For an estimated set of predictors $\breve{\mathcal{P}}_{t+h}$, we first estimate the mutual informations $I(X_{t-\tau}; Y_{t+h})$ for all $X_{t-\tau} \in \breve{\mathcal{P}}_{t+h}$ and choose the one that maximizes the mutual information as our first predictor $X^{(1)}$. Next, we choose the maximal conditional mutual information $I(X_{t-\tau}; Y_{t+h}|X^{(1)})$ among all remaining predictors and obtain $X^{(2)}$. Then $X^{(2)}$ is included in the condition and the iteration converges if all predictors are ranked or some criterion is fulfilled as described below. In each step i, the conditional mutual information gives the gain in information if this predictor is included because

$$\begin{aligned} &I((X^{(1)}, \ldots, X^{(i)}); Y_{t+h}) \\ &= \underbrace{I((X^{(1)}, \ldots, X^{(i-1)}); Y_{t+h})}_{\text{MI without } X^{(i)}} + \underbrace{I(X^{(i)}; Y_{t+h}|(X^{(1)}, \ldots, X^{(i-1)}))}_{\text{gain due to } X^{(i)}}. \end{aligned} \tag{7.5}$$

As a heuristic criterion, we compute the ratio of the information gain of adding predictor $X^{(i)}$ divided by the overall uncertainty reduction of all predictors up to $X^{(i)}$,

$$\lambda_i = \frac{I(X^{(i)}; Y_{t+h}|(X^{(1)}, \ldots, X^{(i-1)}))}{I((X^{(1)}, \ldots, X^{(i)}); Y_{t+h})}, \tag{7.6}$$

where we estimate the denominator by the cumulative sum

$$I((X^{(1)}, \ldots, X^{(i)}); Y_{t+h}) = \sum_{j=1}^{i} I(X^{(j)}; Y_{t+h}|(X^{(1)}, \ldots, X^{(j-1)})), \tag{7.7}$$

and choose as the optimal predictors $\breve{\mathcal{P}}_{t+h}^{(p)} = (X^{(1)}, \ldots, X^{(p)}) \subseteq \breve{\mathcal{P}}_{t+h}$ the maximal number p for which λ_i is still above some predefined threshold λ^*:

$$p = \max\{i : \lambda_i > \lambda^*\}. \tag{7.8}$$

Note that the forward selection need not be globally optimal and an alternative procedure would be to estimate for a given number of predictors i the multivariate MI among all combinations of predictors, which is, however, computationally extremely expensive (Groth, 2001).

The next step is the nearest-neighbor prediction for Y_{t+h}. For the optimal number of predictors p, we first determine the distances

$$d_{t,s} = \|\breve{\mathcal{P}}_{t+h}^{(p)} - \breve{\mathcal{P}}_{s}^{(p)}\| \quad \text{for all} \quad s \in \mathcal{T}, \tag{7.9}$$

where $\mathcal{T}$ is the index set of observed time points $\mathcal{T} = (\tau_{\max} + h_{\max}, \ldots, t)$ and $\|\cdot\|$ denotes some norm, here we use the maximum norm as in the nearest-neighbor estimator of conditional mutual information (Sect. 4.2). For example, if for the prediction step h, we use $\breve{\mathcal{P}}_{t+h}^{(1)} = \{Y_t\}$, we compute all distances from $d_{t,\tau_{\max}+h_{\max}} =$

$\|Y_t - Y_{\tau_{\max}+h_{\max}-h}\|$ to $d_{t,t} = \|Y_t - Y_{t-h}\|$. There are two approaches to use these distances: Either a fixed distance ε is chosen and all points s with distance smaller than ε are used to predict Y_{t+h}, then the coarse-graining level is consistent for all points. Or a fixed number of nearest neighbors is used which has the advantage that the same number of points contribute to a prediction while in the former case often there might not be any point within a distance epsilon (Groth, 2001; Pompe, 2002). We use the latter approach and sort the distances in increasing order $d_{t,s_1} < d_{t,s_2} < \cdots$ yielding an index sequence $s_1, s_2, \ldots$, choose a fixed number of nearest neighbors n and estimate the future value Y_{t+h} by the conditional expectation and its prediction interval by its standard deviation:

$$\widehat{Y}_{t+h} = \frac{1}{n}\sum_{j=1}^{n} Y_{s_j}, \qquad \widehat{\sigma}(\widehat{Y}_{t+h}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(Y_{s_j} - \widehat{Y}_{t+h})^2}. \tag{7.10}$$

Another option, instead of the expectation, is to use an autoregressive model giving a *local-linear prediction* (Farmer and Sidorowich, 1987). The free parameters of this prediction scheme are the parameters of the time series graph estimation as discussed in Sect. 4.4, the number of predictors p taken into account and the number of nearest neighbors n. The estimation of the predictors with the PC algorithm just serves as a pre-selection step and the more crucial parameters for prediction are p and n. The parameter p can be chosen by the heuristic criterion as mentioned above or by cross-validation. The parameter n needs to be balanced to guarantee that local (in phase space) values are used as predictors, but still enough values are available to confidently estimate the mean and variance in Eq. (7.10).

In the linear framework, the first two steps in our prediction scheme are the same, where partial correlation is used instead of conditional mutual information, and in the last step we predict Y_{t+h} using the least squares regression model (3.31) with the predictors $\check{\mathcal{P}}^{(p)}_{t+h} = (X^{(1)}, X^{(2)}, \ldots)$ as regressors for the sample $\mathcal{T}$. Here the prediction interval is given by the variance $\widehat{\sigma}^2_{\varepsilon}$ of the regression residual plus the errors in the estimated regression coefficients $\widehat{\mathbf{B}}$ (Brockwell and Davis, 2002):

$$\widehat{Y}_{t+h} = \check{\mathcal{P}}^{(p)}_{t+h}\widehat{\mathbf{B}}, \qquad \widehat{\sigma}(\widehat{Y}_{t+h}) = \sqrt{\widehat{\sigma}^2_{\varepsilon} + \sum_{i=1}^{p}\widehat{\sigma}(\widehat{\mathbf{B}_i})^2 (X^{(i)})^2}. \tag{7.11}$$

7.2.3. Evaluation of prediction performance

To evaluate the performance of our prediction scheme, we use a *m-fold cross validation* where the set of available observed time indices $\mathcal{T}$ is partitioned into m complementary subsets and for each validation round a subset m is retained as the *testing set* $\mathcal{T}_m$ and the estimation of the time series graph and the ranking of predictors is done on the remaining set $\mathcal{T} \setminus \mathcal{T}_m$. Then the prediction is performed *out-of-sample* on each

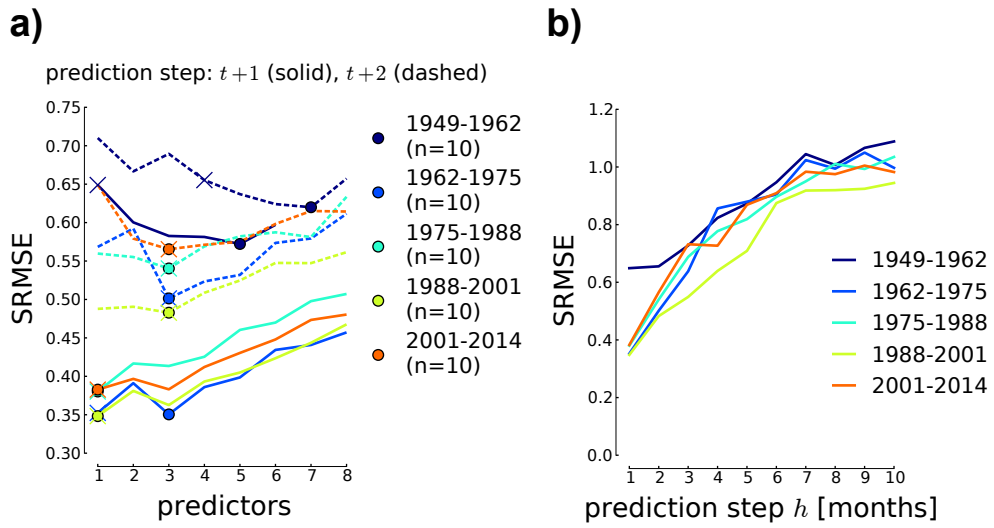


Figure 7.2.: Prediction skill of model-free prediction of the ENSO index Nino3.4 using Eq. (7.10). (a) Standardized root mean squared error given by Eq. (7.12) evaluated using 5-fold cross-validation for the periods indicated in the legend. We show the prediction errors for one (solid) and two months (dashed) ahead predictions plotted against the number of predictors used. The number with minimum out-of-sample error is marked by a colored dot and the optimal number estimated from the learning set with our heuristic criterion (7.8) is marked by a colored cross. $n = 10$ denotes the number of nearest-neighbors used. (b) Prediction error versus the prediction step. For steps larger than 4 months, the error quickly reaches 1, implying that the variance in the prediction is as large as the variance of the respective test period.

testing set and can be evaluated by skill metrics such as the standardized root mean squared error

$$\mathrm{SRMSE} = \sqrt{\frac{\frac{1}{n}\sum_{t\in\mathcal{T}_m}(Y_{t+h} - \widehat{Y}_{t+h})^2}{\widehat{\sigma}_Y^2}}, \tag{7.12}$$

where $\widehat{\sigma}_Y^2$ is the variance of Y in the testing period.

7.3. Predicting El Niño-Southern Oscillation

As an application, we predict the Nino3.4 index defined as the average sea surface temperature over the region (5°N– 5°S, 170°–120°W) using the Hadley SST dataset (Rayner et al., 2003). As predictors, we use the set of 60 sea-level pressure varimax components discussed in Sect. 6.5, but here on a monthly time scale and using the entire season, not just the winter months. We divide the period of January 1948 until February 2014 into five folds of approximately 13 years length each (158 months

for the first four folds and 147 months in the last fold). For each cross-validation round, we leave out one fold and estimate the set of predictors on the remaining folds for a range of prediction steps ahead until $h_{\max} = 10$. For the PC algorithm we use the parameters $\tau_{\max} = 12$, $n_0 = 3$ (initial number of conditions), $n_{\max} = 5$ (maximum number of conditions), and $n_i = 3$ (number of tests per n). The CMI estimation parameter is $k = 100$ and due to the large number of processes ($N = 61$) and because the algorithm is only a pre-selection step, we use a fixed significance threshold $I^* = 0.002$. In the next step, we rank the obtained predictors for each cross-validation round and each prediction step as described above (using the same CMI estimation parameter) and evaluate the prediction on the five test sets using the standardized root mean squared error Eq. (7.12). As a nearest-neighbor parameter we use $n = 10$ to which the results are rather robust (see Appendix B.4).

In Fig. 7.2(a), we show the results for the first two prediction steps. The curves demonstrate that an optimal prediction with minimal out-of-sample error (marked by the colored dots) needs only very few predictors. Especially for $h = 1$, already the first predictor (the Nino3.4 index at the past month for all folds) is optimal (except for the second fold) and including further predictors only leads to overfitting and increases the prediction error on the test set. For $h = 2$ more predictors are needed for an optimal prediction. In most cases our heuristic truncation criterion (7.8) for $\lambda^* = 0.05$ (marked by the colored crosses), which is estimated from the learning set, matches the out-of-sample optimum.

For example, in the last fold 2001–2014 for the prediction of Nino3.4$(t + 2)$ the first predictor is Nino3.4 at time t which reduces the uncertainty (given by the MI) by 0.59 nats in the learning set. On the test set the prediction error here is 0.65. The next ranked predictor is the pressure component No. 1 at the same lag in the East Pacific region (see map in Fig. 6.7) which further reduces the uncertainty by 0.03 nats and decreases the error to 0.58. The third predictor No. 0 in the Eastern Indian Ocean slightly further decreases the uncertainty by 6% yielding a minimum prediction error of 0.57 after which the next predictor only reduces the uncertainty by 1% and the prediction error now steadily increases.

The first step prediction error is quite comparable among the folds (with the exception of the first fold), while for larger steps the prediction errors vary more. In Fig. 7.2(b), we show the errors plotted against the prediction step using the predictors chosen by our heuristic criterion. For steps larger than 4 months, the error quickly reaches 1 which implies that the prediction error is as high as the variance of the test data and thus, the prediction is merely a persistence forecast.

If the same approach is used in the linear framework with a prediction using Eq. (7.11), the prediction skill is considerably improved, especially in the short term, as shown in Fig. 7.3(b), but also here after more than 6 months the skill is rather low. The better linear prediction is a sign that Nino3.4 can be well modeled by a linear process on these short time scales. In the linear case the problem of overfitting is also not that severe and the prediction error is not increasing much with the number of predictors (Fig. 7.3(a)). Still, the number of predictors giving minimal out-of-sample error is usually rather small in the range of 5–10 here which is also the range given

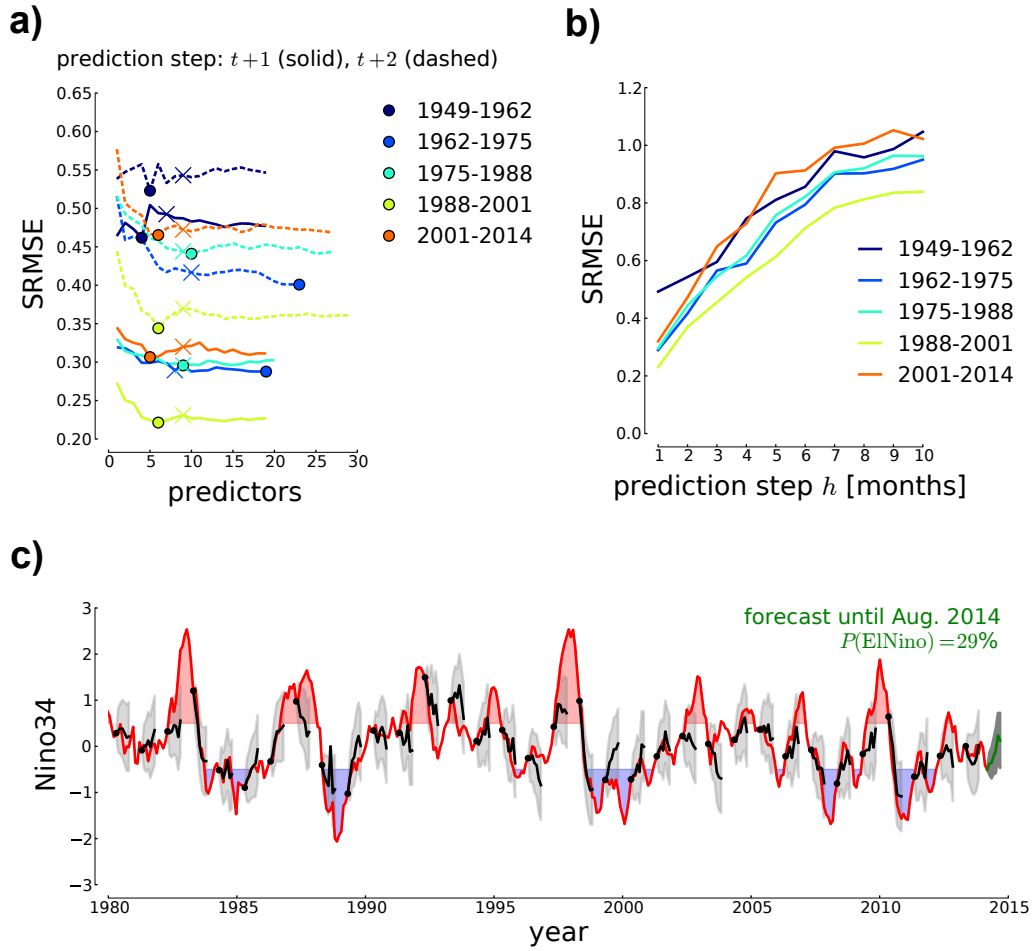


Figure 7.3.: Prediction skill of linear prediction of ENSO according to Eq. (7.11). (a) and (b) are as in Fig. 7.2, in (c) we show the Nino3.4 index with El Niño and La Niña events marked in red and blue, respectively. The black lines denote selected hindcasts and their prediction intervals (grey) starting from April in each year. The green line marks a real forecast starting from February 2014 giving a probability for an El Niño event of about 30% until August. Steps further in the future have higher probability, but are not shown since the prediction skill for the past decade is too low to be confident about this forecast.

by our heuristic criterion. To give an impression of selected predictions (actually hindcasts), we show in Fig. 7.3(c) the predictions up to 6 months starting from April for the period 1980–2014. The important onsets of El Niño events (defined as the Nino3.4 index above 0.5 °C, while below −0.5 °C are La Niña events) are partially predicted (e.g., for the 1982, 1987 and 1997 events), but especially in the last decade, almost no El Niño event is predicted even only 4 months ahead and one prediction even tends in the wrong direction (event 2003). This bad predictability of the recent

decade is also found in other studies (Barnston et al., 2012) and suggests that the mechanism of ENSO is changing.

Finally, we give a real forecast using those predictors and n to minimize error in the last fold. The green line in Fig. 7.3(c), starting from February 2014, shows the forecast. The probability of an El Niño event, computed as the probability that the forecast distribution exceeds 0.5 °C (assuming a Gaussian distribution with mean and standard deviation given by Eq. (7.11)), is maximal for August 2014 with an about 30% chance. The probability rises in later months (which is also the usual El Niño season), however, given the low skill in the last decade, we have not much confidence into this forecast. The particular problem of forecasting the ENSO season more than a few months ahead has been termed the "spring barrier" (Webster and Yang, 1992; Webster and Hoyos, 2010). While we used a scheme to predict the value of a time series, one can also formulate a different goal and forecast events as done in Ludescher et al. (2013); Ludescher et al. (2014) where an El Niño event for 2014 is predicted with 75% likelihood even more than a year ahead.

7.4. Summary – a model-free baseline for prediction

In this chapter, we have shown how the causal inference algorithm can be used to obtain optimal model-free predictors. Optimal in the sense that they provide a maximum of information about the variable to be predicted. Since this might be of little use due to the curse of dimensionality in a realistic situation given finite samples, we propose a ranking scheme which allows to evaluate how much gain in information the addition of a predictor gives. Too few predictors lack useful information for prediction, but each added predictor increases the dimensionality and raises the chance of overfitting, i.e., the model adapts to the noise rather than the process. Our ranking procedure combined with a heuristic truncation criterion allows to balance out these two effects as we have shown in an application to forecast the Nino3.4 index. One might question whether the causal pre-selection step improves a prediction compared to using forward selection alone. Tillman and Spirtes (2008) compared causal prediction methods (without further ranking and truncation against overfitting) to pure prediction methods that do not take into account the conditional dependencies but avoid overfitting. They found that purely causal methods tended to give worse predictions due to overfitting, while pure prediction methods failed in some cases if their selection procedure did not include causal predictors, which is also consistent with our analytical derivation of optimality (Eq. (7.3)). They conclude that an optimal prediction method should be based on causal predictors combined with methods avoiding overfitting, which is exactly what we propose. But whether our causal pre-selection step *strongly* improves a prediction compared to directly using the forward selection scheme will depend on the specific process.

The prediction scheme can be further optimized. One way of improvement lies in weighting the predictors differently, for example, according to their distance to $\breve{\mathcal{P}}_{t+h}^{(p)}$. This also includes the question of which distance metric (e.g., a data-adaptive Mahalanobis distance metric which scales according to the multivariate covariance of the data) is most suitable for a given problem. Here we kept the estimation scheme simple, but these further free parameters could considerably improve a prediction. In general, such hyper-parameters can be optimized by cross-validation as well[13]. In our scheme we used a symmetric prediction interval constructed from the standard deviation, but the conditional prediction distribution could also be asymmetric, possibly important to predict extreme events. Also, the prediction can be separately studied for different seasons using non-stationary time series graphs (Sect. 2.4.5).

While we use a nearest-neighbor prediction scheme for the model-free case, also here the framework can be applied using linear measures. One can view the model-free prediction as a lower baseline against which a model-based prediction can be evaluated. In our case the linear approach considerably improved the prediction. A model-based prediction that does not improve a model-free prediction implies that the model assumptions invoked are possibly misspecified.

[13]Note that such an optimization must be strictly separated from using cross-validation to evaluate the performance of a prediction scheme.

Chapter 8.

Conclusion

We can't solve problems by using the same kind of thinking we used when we created them.

A. Einstein

8.1. A multidisciplinary feedback loop

Often the pursuit of a research question leads to problems that are similarly encountered in other disciplines. This is especially true for the research on complex systems from the perspective of time series analysis and even more so for the fundamental question of inferring causality from time series. As said in the beginning, I deem the combination and exchange of different perspectives from various disciplines to be the best approach to understand complex systems. Even though it was time-consuming to learn a different language of science, the fruitfulness of this approach has been shown in many rewarding encounters in the course of this doctoral research.

I hope that the methods developed in this thesis as summarized below are not only useful for the physical perspective on interactions, but that some aspects are also valuable for statistics and machine learning. But ultimately, I hope that my methods can contribute a small piece in understanding great challenges of humankind such as climate change in the spirit that science should not only *"open the door to infinite wisdom, but set some limit on infinite error"* (B. Brecht).

8.2. Contributions of this thesis and outlook

In the following, the contributions to the fields of nonlinear time series analysis of complex systems, statistics, and climate research that have been reported in this thesis are summarized together with an outlook to promising avenues of further research. For reference, the corresponding sections are given and relevant publications (co-)authored by myself are listed below the respective paragraphs.

8.2.1. Nonlinear time series analysis of complex systems and information theory

Detecting causality

My co-authors and I have addressed the problem of inferring causal relations among multiple processes within the model-free information-theoretic framework utilizing the concept of time series graphs as a generalization of model-based Granger causality. The modified algorithm to iteratively determine such a causal network of a multivariate process constitutes a novel contribution to the physics of complex systems (Sections 2.4.6 and 4.4) and we were also the first to apply and numerically study it together with the recently introduced estimator of conditional mutual information (Sect. 4.4). This approach alleviates the curse of dimensionality and allows for an application to multivariate time series as shown in several examples and numerical experiments with a very general class of nonlinear stochastic processes (Sect. 4.4.3). The concept of time series graphs also includes precise coupling delays and contemporaneous links. Numerical studies of the estimator of conditional mutual information (Chapter 4, summarized in Table 4.1) also demonstrated the limitations of such a model-free approach. Regarding the question of how many processes can maximally be taken into account, we emphasize that due to the efficient iterative testing scheme of the PC algorithm, the limiting factor in the reconstruction of causal networks from multivariate time series data is not the network size, but only the maximum 'true' degree, i.e., the number of causal parents, while the number of processes can be much higher if they are sparsely connected. One important problem often overseen in assessing the significance of associations are autocorrelations for which measures such as mutual information, but also the more sophisticated transfer entropy, tend to a much higher rate of false detections (Sect. 4.3.3). With the novel measure *momentary information transfer* (MIT), this problem is largely overcome, providing more reliable conditional independence tests that are not biased towards autocorrelated time series (Sect. 4.3.3 and Table 4.1). The main focus of application in this thesis were climatological time series (Chapter 6), where the problem of autocorrelation is especially severe. The insights of these applications are discussed below in the section on climate science.

The model-free causal inference approach introduced here is applicable also to other fields of science. For example, in a field such as neuroscience where much less is known about the interplay between different regions in the brain. The requirements, next to the fundamental assumptions of causal inference (Sect. 2.4.7), are that the time series attain a continuous range of values for which the nearest-neighbor estimator used here is applicable and they need to be sufficiently long and stationary to obtain reliable estimates. One particular non-stationarity occurring especially in climate time series is a seasonally changing causality that can be addressed by a modified concept of time series graphs (Sect. 2.4.5). For non-continuous variables, such as event time series, the discrete "plug-in"-estimators of information-theoretic quantities have to be used which do not allow to exploit metrics as discussed in Sect. 4.5.

On the technical and conceptual side, several improvements are possible. Of particular importance is the development of better estimators of conditional mutual information that reduce bias and variance. This can probably be achieved by relaxing the idea of "total model-freeness," and constructing estimators that incorporate assumptions reasonable for a specific application. Also the theoretical properties of estimators need to be further studied. At present, no results on the sample distribution or even its moments are known which would help in assessing significance levels analytically, rather than with computationally expensive shuffle tests. On a conceptual side, here we defined time series graphs with nodes given by every subcomponent of a process at a certain time. But one can argue from the perspective of dynamical systems theory, that for some systems not these measurements are of interest as nodes, but *states* reconstructed by delay embedding (Kantz and Schreiber, 2003), and our approach could be generalized by defining the time series graph of states.

Quantifying interactions

Our second research aim was to quantify complex interactions in a well-interpretable way. For the quantification of a link in a causal network, we have shown analytically and numerically that the commonly used measures mutual information and transfer entropy can be rather unintuitive as measures of coupling strength (Sect. 5.2.2) and transfer entropy, further, also suffer from estimation bias (Sect. 5.4). To overcome this limitation, we introduced the information-theoretic MIT based on the physically motivated concept of source entropy (Sect. 3.4.5). MIT fulfills a set of proposed properties (Sect. 3.1.2, summarized in Tab. 3.1) that allow for an intuitive interpretation. One property, coupling strength autonomy, allows to disentangle the different factors constituting an interaction mechanism, as we prove analytically in a theorem and numerically (Sections 5.3 and 5.4). We find that the coupling strength autonomy property is useful mostly for models of processes where the coupling strength can be attributed to one single coefficient, while for other cases we suggest modifications of momentary information transfer as more appropriate measures (Sections 3.4.4 and 3.5.3). For transfer entropy, we also derive an exact decomposition formula that enables an estimation using finite vectors (Sect. 3.4.2). In an excursional section, we provided relations of our approach to communication theory (Sect. 5.5.1), an attempt of a thermodynamic perspective (Sect. 5.5.2) and links to geophysical processes (Sect. 5.5.3). The advantage of MIT for climatological interpretations of causal links is studied in Sections 6.3 and 6.4.

We extended the idea of momentary information also to more complex interaction schemes involving multiple processes along causal paths (Sections 3.5.1) and introduced measures that allow to determine and quantify which intermediate processes are mediating such an influence (Sect. 3.5.2). These measures were studied on analytical examples (Sections 5.2.3 and 5.2.4) and we also discussed how these measures can be used to quantify more global properties of information transfer complementing complex network theory (Sect. 5.5.4). In Appendix A.5 the results on interaction

information are further expanded providing a contribution to information-theoretic aspects of Gaussian channels. The applications in Sect. 6.5 demonstrate the potential of such measures in climate research and beyond.

As two more potential uses of our approach, we discussed the detection of critical transitions and time series prediction. For the former, we studied how MIT, estimated over time in sliding windows, can serve as a measure to determine critical transitions in the strength of causal mechanisms in a preliminary example of the interaction between ENSO and the Indian Summer Monsoon (Sect. 6.6). The knowledge of causal drivers can also help in optimally predicting time series in a model-free way constituting a baseline of what can be predicted with as few assumptions as possible from data alone (Chapter 7).

The limitations of estimating conditional mutual information also affect an assessment of a causal strength, especially since the increasing bias of the CMI estimator in higher dimensions makes it difficult to compare MIT values estimated with conditions of different dimension. Such a task demands even longer time series limiting the range of applicability of model-free methods. This is also the reason why we developed the framework in parallel using linear measures that can be much better estimated even for very short time series. The intuitive interpretability also carries over to linear measures (Sect. A.6). While we have argued here that the idea of momentary information provides an intuitive measure of causal strength, other authors may find other approaches more intuitive such as the idea in Janzing et al. (2013) inspired by communication theory (Sect. 5.5.1). In applications, the practitioner will have to decide which measure best suits her demands.

The causal view of interactions in complex systems opens a number of avenues for further research. The knowledge of the causal parents, paths, and motifs in the network can be used to study general forms of information processing in complex systems (Milo et al., 2002), to define measures that quantify distinct aspects such as the interactions between sets of nodes or to predict how perturbations propagate in a complex system. On the side of theoretical applications we have mainly studied nonlinear stochastic processes, but the detailed, probably numerically based, study for coupled chaotic systems might yield deep insights into information flow in these systems following works by Liang (2013).

Related publications Runge et al. (2012a); Runge et al. (2012b); Hlinka et al. (2013); Balasis et al. (2013)

8.2.2. Statistics and machine learning

Apart from learning from statistics, our model-free causal inference approach combining the PC algorithm with recent estimators of conditional mutual information (Sect. 4.4) also constitutes a novel approach for applied statistics and machine learning. The causal inference framework can equally be invoked with linear measures of causality which can be much better estimated (Table 4.1). For classical linear statistics, our study of the partial correlation MIT estimator in the presence of autocorrelation

(Sect. 4.3.2) provides a possible solution overcoming the violation of the assumption of independent and identically distributed samples, that enables the use of analytical theory for significance testing. Also our study on the relation between the partial correlation MIT and autoregressive processes constitutes an original contribution to classical time series analysis (Sect. A.6). Finally, our optimal model-free prediction approach (Sect. 7) is a useful contribution to this core topic of machine learning.

Autocorrelations are an ubiquitous feature of time series not only in climate science, but also in many other fields such as economics and neuroscience, and our approach to overcome autocorrelation in the detection as well as the quantification of associations could be utilized as an elegant alternative to methods such as *pre-whitening* or *first differencing* (Sect. 4.3) that are common in economics. This constitutes an example of using conditioning to eliminate autocorrelation as a nuisance parameter in the framework of conditional inference (Neyman and Scott, 1948; Reid, 1995; Amarasingham et al., 2012) which effectively increases the sample size. The latter approach can also be further developed, for example, by conditioning an association measure between two time series on a signal with a much slower time scale underlying both of them.

Related publications Runge et al. (2012a); Runge (2013)

8.2.3. Climate research

The main focus of application has been the study of climate interactions from the 20th century up to now, in particular in the tropics connected to El Niño-Southern Oscillation (ENSO) as a main driver of global climate. As a methodological contribution relevant not only in climate research, we have studied the pitfalls of using the cross correlation lag function and regressions to assess possible time delays and the strength of a climatic mechanism in the presence of autocorrelations (Sect. 5.2.1). The conclusion is that these measures are quite ambiguously influenced by such internal dynamics with strong inertia (e.g., a large oceanic heat capacity) and misguide an estimate of a physical coupling strength and delay. The same conclusions also hold for lag functions of other measures like mutual information if the effect of autodependencies is not conditioned out. These problems can be overcome using the causal inference algorithm and the subsequent estimation of causal strength, which we utilized here with nonlinear and linear measures (Sect. 2.4).

On the side of climatological results, we detected that the coupling mechanism of the tropical East Pacific on the northern tropical Atlantic has a short lag of one month consistent with the advection speed of the Pacific – Atlantic Walker circulation, while previous studies using the maximum of the cross correlation lag function found lags of 3–6 months (Sections 6.3 and B.2). Also, we uncovered that the coupling mechanism is actually quite weak (even comparable to the coupling mechanism between Western and Eastern Europe) and that the large cross correlation value can be explained by strong autocorrelations present in both time series. As a further step, we demonstrated the potential of our approach to identify interaction

mechanisms also between more than two processes by investigating the mechanism of the Walker circulation (Sect. 6.4). The purely statistical analysis confirmed that the positive correlation between surface temperatures over the Eastern Pacific and surface pressure over the Western Pacific is mediated via the Central Pacific while the lagged correlation back cannot be explained by variabilities in surface temperatures of the Central Pacific. These examples validate the potential use of our methods to test specific hypotheses of complex interactions in climate research and beyond.

In a more exploratory analysis, we studied a novel technique of analyzing causal interactions in larger climatic fields, here a global surface pressure dataset (Sect. 6.5). First, we reduced the dimensionality of the more than $10,000$ time series to a limited set accounting for well interpretable and partially well-known subprocesses like ENSO. The analysis of causal interactions among these components with more aggregated measures revealed that many tropical processes strongly influence causally adjacent nodes, but only ENSO's influence can be strongly measured also in nodes further apart in the network. On the other hand, we found that in other characteristics such as the 'causal interaction betweenness' other processes such as in the East Indian Ocean are even stronger in impacting on the interactions between pairs in the network (Sect. 6.5.4). A more detailed analysis of selected interaction pathways demonstrated the usefulness of the interaction measure approach to determine the most probable causal paths on which a climatic mechanism is mediated. For example, consistent with known climatological processes we found that the impact of ENSO on West Australia is mediated via the East Indian Ocean (Sect. 6.5.5).

Further, we studied the recent abrupt weakening of the relationship between ENSO and the Indian Monsoon as a possible tipping point for which our novel measures might serve as early warning signals (Sect. 6.6). Finally, we showed how causal interactions could be used as optimal predictors, here applied to the statistical prediction of ENSO (Sect. 7.3). In an appendix, we give more examples of Pacific – Atlantic interactions (Sect. B.2) and the coupling between surface and tropospheric temperatures in the tropics (Sect. B.3). In Schleussner et al. (2014), we studied another application of linear theory in order to causally decompose the contributions of the Atlantic meridional overturning circulation (AMOC) and Arctic sea ice to global mean temperature. We found that a significant contribution of the AMOC to global mean temperature stems from the AMOC feedback with sea-ice.

One limitation for the use of information-theoretic methods is the short length of climate time series. Of many climatological variables reliable measurements exist only since the 1980s where satellite observations began. Further, we discussed in Sect. 6.7 that on the considered monthly and weekly time scales often interactions can be assumed to be linear. Therefore, information-theoretic methods might be of more use on shorter daily weather time scales where nonlinearities together with causal methods might also be key to extend the prediction horizon presently limited to about 10 days.

The application of causal measures in climate research – introduced in this work after Ebert-Uphoff and Deng (2012b) – is still in its infancy and many more interesting climatological questions can be addressed in the hypothesis-driven application of our

two-step approach ranging from determining mechanisms driving the Indian Monsoon to understanding ENSO's influence on global climate. Also in an exploratory way, the interactions in and between different climatological variables can be studied from a complex systems perspective complementing and advancing the approaches by Tsonis and Roebber (2004); Donges et al. (2009a).

Related publications Runge et al. (2012b); Schleussner et al. (2014); Runge et al. (2014)

Appendix

Appendix A.

Analytical derivations, proofs, and further theoretical results

A.1. Derivation of decomposed transfer entropy

This section will give formal derivations and proofs for Eqns. (3.40), (3.41) and (3.42) (Runge et al., 2012a). For Eq. (3.40), the chain rule of conditional mutual information (Cover and Thomas, 2006) is iteratively applied to TE:

$$\begin{aligned} I^{\mathrm{TE}}_{X\to Y} &= I(X_t^-; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-) \\ &= I(X_{t-1}^-; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-) + I(X_{t-1}; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-, X_{t-1}^-) \\ &\vdots \\ &= \underbrace{\lim_{\tau\to\infty} I(X_{t-\tau}^-; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-)}_{=0 \quad \text{if } \tau^{\max}_{X\to Y} \text{ is finite}} + \sum_{\tau=1}^{\infty} I(X_{t-\tau}; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-, X_{t-\tau}^-), \end{aligned} \tag{A.1}$$

where the first term $I(X_{t-\tau}^-; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-)$ is zero provided we have a finite maximum coupling delay $\tau^{\max}_{X\to Y}$.

Equation (3.41),

$$I(X_{t-\tau}; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^-, X_{t-\tau}^-) = I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}), \tag{A.2}$$

is derived by applying the chain rule to a certain multivariate CMI in two different ways:

$$\begin{aligned} &I(X_{t-\tau}, \{\mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-\} \backslash \mathcal{S}_{Y_t, X_{t-\tau}}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}) = \\ &= I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}) + \underbrace{I(\{\mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-\} \backslash \mathcal{S}_{Y_t, X_{t-\tau}}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}, X_{t-\tau})}_{\text{Select } \mathcal{S}_{Y_t, X_{t-\tau}} \text{ such that this term} = 0} \\ &= \underbrace{I(\{\mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-\} \backslash \mathcal{S}_{Y_t, X_{t-\tau}}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}})}_{\text{Select } \mathcal{S}_{Y_t, X_{t-\tau}} \text{ such that this term} = 0} + I(X_{t-\tau}; Y_t \,|\, \mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-)\,, \end{aligned} \tag{A.3}$$

where now the set of nodes $\mathcal{S}_{Y_t, X_{t-\tau}}$ has to be chosen such that it separates the set of infinite conditions $\{\mathbf{X}_t^- \backslash X_t^- \cup X_{t-\tau}^-\} \backslash \mathcal{S}_{Y_t, X_{t-\tau}}$ from Y_t in the graph (Eq. (2.12).

We therefore arrive at a finite set of conditions:

$$I^{\mathrm{TE}}_{X\to Y} = \sum_{\tau=1}^{\infty} I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}). \tag{A.4}$$

The last remaining infiniteness now lies in the infinite sum, but – as in the derivation of Eq. (3.40) – we argue that under mild assumptions the summands should decay exponentially, just like in the case of a vector-autoregressive process (Brockwell and Davis, 2009). Then one can truncate the sum at some $\tau^\star$ leading to Eq. (3.42) for the

decomposed transfer entropy (DTE)

$$I^{\mathrm{TE}}_{X\to Y} \approx I^{\mathrm{DTE}}_{X\to Y} = \sum_{\tau=1}^{\tau^\star} I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}) \,. \tag{A.5}$$

One can also approximate the rest of the sum by several of the last terms for a more precise estimator of TE,

$$I^{\mathrm{TE}}_{X\to Y} \approx \sum_{\tau=1}^{\tau^\star-1} I(X_{t-\tau}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau}}) + \frac{I(X_{t-\tau^\star}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau^\star}})}{1-\delta} \,, \tag{A.6}$$

with $\tau^\star$ chosen as the smallest τ for which the estimated remainder is smaller than some given absolute tolerance, and a damping factor δ estimated from several of the last terms, e.g.,

$$\delta = \frac{I(X_{t-\tau^\star-1}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau^\star-1}})}{I(X_{t-\tau^\star}; Y_t \,|\, \mathcal{S}_{Y_t, X_{t-\tau^\star}})}. \tag{A.7}$$

A.2. Derivations of correlation lag function and regressions for model Eq. (5.1)

Here, we derive the analytical expressions for the (co-)variances needed to evaluate the regressions, cross correlation and the partial correlations ITY and MIT shown in Tab. 5.1. The model Eq. (5.1), here better discussed in the form of Eq. (2.17),

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \underbrace{\begin{pmatrix} a & 0 \\ c & b \end{pmatrix}}_{\Phi(1)} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{X,t} \\ \varepsilon_{Y,t} \end{pmatrix} \,. \tag{A.8}$$

belongs to the general class of vector autoregressive processes of order p defined as

$$\mathbf{X}_t = \sum_{s=1}^{p} \Phi(s) \mathbf{X}_{t-s} + \varepsilon_t, \tag{A.9}$$

where $\Phi(s)$ are the $N \times N$ matrices of coefficients for each lag s and the N-vector $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$ is an independent identically distributed Gaussian random variable with zero mean and covariance matrix Σ. ε is sometimes referred to as the *innovation term*. Its variances on the main diagonal of Σ are denoted by σ_i^2 and the covariances by σ_{ij}.

For this model, there exists an analytical expression of the covariance in terms of Φ (Brockwell and Davis, 2009, Ch. 11.3):

$$\Gamma_{ij}(\tau) \equiv E[\mathbf{X}^i_{t+\tau} \mathbf{X}^j_t] = \sum_{n=0}^{\infty} \Big(\Psi(n+\tau) \Sigma \Psi^\top(n) \Big)_{ij} \tag{A.10}$$

where the matrix $\Psi(n)$ can be recursively computed from matrix products:

$$\Psi(n) = \sum_{s=1}^{n} \Phi(s)\Psi(n-s). \tag{A.11}$$

In the case of an autoregressive model of first order model, all coefficient matrices $\Phi(s)$ with lags $s > 1$ are zero, and as can be seen from Eq. (A.11), Ψ is simply given by the matrix powers of $\Phi(1)$ which are

$$\Psi(n) = \Phi(1)^n = \begin{pmatrix} a & 0 \\ c & b \end{pmatrix}^n = \begin{pmatrix} a^n & 0 \\ (a^n - b^n)\frac{c}{a-b} & b^n \end{pmatrix}. \tag{A.12}$$

Then the variances for $\tau = 0$ are

$$\begin{aligned} \Gamma_X &= \sum_{n=0}^{\infty} \left(\Psi(n)\Sigma\Psi^\top(n)\right)_{XX} \\ &= \sigma_X^2 \sum_{n=0}^{\infty} a^{2n}, \end{aligned} \tag{A.13}$$

$$\begin{aligned} \Gamma_Y &= \sum_{n=0}^{\infty} \left(\Psi(n)\Sigma\Psi^\top(n)\right)_{YY} \\ &= \frac{1}{(a-b)^2} \sum_{n=0}^{\infty} \Big[\Big((a-b)^2\sigma_Y^2 + c\left(c\sigma_X^2 - 2a\sigma_{XY} + 2b\sigma_{XY}\right)\Big) b^{2n} + \\ &\qquad + a^{2n}c^2\sigma_X^2 - 2a^n c\left(c\sigma_X^2 - a\sigma_{XY} + b\sigma_{XY}\right) b^n\Big]. \end{aligned} \tag{A.14}$$

Noting that the infinite sums are geometric series that converge assuming $0 < |a|, |b| < 1$, one arrives at the variances in Tab. 5.1 (where additionally σ_{XY} was set to 0). Similary, the covariance function for the direction $Y \to X$ (valid for $\tau \leq 0$) is

$$\begin{aligned} \Gamma_{XY}(\tau) &= \sum_{n=0}^{\infty} \left(\Psi(n+\tau)\Sigma\Psi^\top(n)\right)_{XY} \\ &= \frac{1}{a-b} \sum_{n=0}^{\infty} a^{n+\tau}\left(a^n c\sigma_X^2 - b^n\left(c\sigma_X^2 - a\sigma_{XY} + b\sigma_{XY}\right)\right), \end{aligned} \tag{A.15}$$

and for the direction $X \to Y$ (valid for $\tau > 0$) is

$$\begin{aligned} \Gamma_{YX}(\tau) &= \sum_{n=0}^{\infty} \left(\Psi(n+\tau)\Sigma\Psi^\top(n)\right)_{YX} \\ &= \frac{1}{a-b} \sum_{n=0}^{\infty} a^{n}\left(a^{n+\tau} c\sigma_X^2 - b^{n+\tau}\left(c\sigma_X^2 - a\sigma_{XY} + b\sigma_{XY}\right)\right), \end{aligned} \tag{A.16}$$

from which the cross correlation in Tab. 5.1 follows (with $\sigma_{XY} = 0$). Note that $\Gamma_{YX}(\tau)$ does not diverge for $a = b$ since in this limit according to L'Hôpital's rule:

$$\Gamma_{YX}(\tau) \overset{a=b}{=} \frac{b^{\tau-1}\left(bd\left(1-b^2\right) - b^2 c\sigma_X^2(\tau-1) + c\sigma_X^2\tau\right)}{\left(1-b^2\right)^2}. \tag{A.17}$$

As a check, for no autocorrelation, i.e., for $a = b = 0$ and at the correct coupling lag $\tau = 1$, this gives

$$\Gamma_{YX}(1) \overset{a=b=0}{=} c\sigma_X^2. \tag{A.18}$$

The inequality relation in the caption of Fig. 5.1 for zero contemporaneous dependency $\sigma_{XY} = 0$ is obtained from simplifying $\Gamma_{YX}(2) > \Gamma_{YX}(1)$ using the assumption that a, b are positive and smaller than 1. The regression coefficients are gained by inserting the previously derived covariances into the regression formula in Eq. (3.34). ITY can be derived by analogously computing $\Gamma_{YY}(1)$ and using the fact that the partial correlation $\rho(X_{t-1}; Y_t|Y_{t-1})$ is equivalent to the cross correlation of the residuals of X_{t-1} and Y_t after regression on Y_{t-1}. This leads to the residual covariance $\Gamma_{YX}(1) - \Gamma_{YX}(0)\Gamma_{YY}(1)/\Gamma_Y$ and the residual variances $\Gamma_Y - \Gamma_{YY}(1)^2/\Gamma_Y$ and $\Gamma_X - \Gamma_{YX}(0)^2/\Gamma_Y$ from which the value in Tab. 5.1 follows (with $\sigma_{XY} = 0$). MIT could be similarly computed, but also follows from the linear version of the coupling strength autonomy theorem given in Sect. A.6.2.

A.3. Derivations for model Eq. (5.3)

In this section, we derive TE, MI, MIT and related measures for the coupling between X and Y in model Eq. (5.3),

$$\begin{aligned} Z_t &= c_{XZ}X_{t-1} + \eta_t^Z \\ X_t &= a_X X_{t-1} + \eta_t^X \\ Y_t &= c_{XY}X_{t-2} + c_{WY}W_{t-1} + \eta_t^Y \\ W_t &= \eta_t^W \end{aligned} \tag{A.19}$$

with independent Gaussian white noise processes $\eta_t^\cdot$ with variances $\sigma_\cdot^2$.

A.3.1. Derivation for TE

For the derivation of TE

$$I_{X\to Y}^{\mathrm{TE}} = H(Y_t|Y_t^-, W_t^-, Z_t^-) - H(Y_t|X_t^- Y_t^-, W_t^-, Z_t^-), \tag{A.20}$$

we know from the Markov property that the latter term is the source entropy

$$H(Y_t|\mathcal{P}_{Y_t}) = \frac{1}{2}\ln 2\pi e\sigma_Y^2\,. \tag{A.21}$$

For the first entropy

$$H(Y_t|Y_t^-, W_t^-, Z_t^-) = \frac{1}{2}\ln\left(2\pi e\frac{|\Gamma_{Y_tY_t^-W_t^-Z_t^-}|}{|\Gamma_{Y_t^-W_t^-Z_t^-}|}\right), \tag{A.22}$$

we can write the covariance as a block matrix

$$\Gamma_{Y_tY_t^-W_t^-Z_t^-} = \begin{pmatrix} \Gamma_{Y_t} & \Gamma_{Y_t;Y_t^-} & \Gamma_{Y_t;W_t^-} & \Gamma_{Y_t;Z_t^-} \\ \Gamma_{Y_t;Y_t^-}^\top & \Gamma_{Y_t^-} & \Gamma_{Y_t^-;W_t^-} & \Gamma_{Y_t^-;Z_t^-} \\ \Gamma_{Y_t;W_t^-}^\top & \Gamma_{Y_t^-;W_t^-}^\top & \Gamma_{W_t^-} & \Gamma_{W_t^-;Z_t^-} \\ \Gamma_{Y_t;Z_t^-}^\top & \Gamma_{Y_t^-;Z_t^-}^\top & \Gamma_{W_t^-;Z_t^-}^\top & \Gamma_{Z_t^-} \end{pmatrix}. \tag{A.23}$$

where, e.g., $\Gamma_{Y_t;W_t^-}$ is an infinite vector with entries of the covariances of Y_t with W_{t-1}, $W_{t-2}, \ldots$ and

$$\Gamma_{Y_t^-;W_t^-} = \begin{pmatrix} \Gamma_{YW}(0) & \Gamma_{YW}(1) & \ldots \\ \Gamma_{WY}(1) & \Gamma_{YW}(0) & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

The quotient in Eq. (A.22) of the determinants of these infinite dimensional matrices is difficult, if not impossible, to evaluate in the general case. Here, we will only consider two cases.

Case $c_{XZ} = c_{WY} = 0$

For the case of $c_{XZ} = c_{WY} = 0$, i.e., as inputs solely an autodependency in X, the covariance matrix takes the simple form

$$\Gamma_{Y_tY_t^-W_t^-Z_t^-} = \begin{pmatrix} \Gamma_{Y_t} & \Gamma_{Y_t;Y_t^-} & 0 & 0 \\ \Gamma_{Y_t;Y_t^-}^\top & \Gamma_{Y_t^-} & 0 & 0 \\ 0 & 0 & \Gamma_{W_t^-} & 0 \\ 0 & 0 & 0 & \Gamma_{Z_t^-} \end{pmatrix} \tag{A.24}$$

where the top left block is an infinite dimensional Toeplitz matrix, i.e., a Toeplitz operator. Then the quotient in Eq. (A.22) can be simplified to

$$\frac{|\Gamma_{Y_tY_t^-}||\Gamma_{W_t^-Z_t^-}|}{|\Gamma_{Y_t^-}||\Gamma_{W_t^-Z_t^-}|} = \frac{|\Gamma_{Y_tY_t^-}|}{|\Gamma_{Y_t^-}|}. \tag{A.25}$$

$\Gamma_{Y_t Y_t^-}$ and $\Gamma_{Y_t^-}$ are the symmetric Toeplitz matrices G_τ and $G_{\tau-1}$ with diagonal elements Γ_Y and off-diagonal elements g_τ

$$g_0 = \Gamma_Y = c_{XY}^2 \frac{\sigma_X^2}{1-a_X^2} + \sigma_Y^2 \tag{A.26}$$

$$g_\tau = a_X^{|\tau|} \frac{c_{XY}^2 \sigma_X^2}{1-a_X^2} \qquad \text{for } \tau \geq 1. \tag{A.27}$$

The desired TE is then given by

$$I_{X\to Y}^{\mathrm{TE}} = \lim_{\tau\to\infty} \frac{1}{2} \ln \frac{1}{\sigma_Y^2} \frac{|G_\tau|}{|G_{\tau-1}|}. \tag{A.28}$$

To obtain the limit of the ratio of infinite Toeplitz matrix determinants, we can utilize Szegö's theorem (Szegö, 1915; Böttcher et al., 2006) which relates the limit to the geometric mean of a function $f(\lambda)$

$$\lim_{\tau\to\infty} \frac{|G_\tau(f)|}{|G_{\tau-1}(f)|} = \exp\left(\frac{1}{2\pi}\int_0^{2\pi} \ln f(\lambda) d\lambda\right), \tag{A.29}$$

which requires that the Toeplitz matrix is in the Wiener class, i.e., the entries must be absolutely summable, which we assume here. The function $f(\lambda)$ is the Fourier series with the entries of the Toeplitz matrix being the coefficients

$$f(\lambda) = \sum_{\tau=-\infty}^{\infty} g_\tau e^{i\tau\lambda} = \Gamma_Y + 2\sum_{\tau=1}^{\infty} g_\tau e^{i\tau\lambda} \tag{A.30}$$

$$= c_{XY}^2 \frac{\sigma_X^2}{1-a_X^2} + \sigma_Y^2 + 2\frac{c_{XY}^2\sigma_X^2}{1-a_X^2} \underbrace{\sum_{\tau=1}^{\infty} a_X^{|\tau|} e^{i\tau\lambda}}_{\frac{a_X e^{i\lambda}}{1-a_X e^{i\lambda}}} \tag{A.31}$$

$$= \frac{\overbrace{[c_{XY}^2\sigma_X^2 - \sigma_Y^2(1-a_X^2)]a_X}^{\alpha} e^{i\lambda} + \overbrace{c_{XY}^2\sigma_X^2 + \sigma_Y^2(1-a_X^2)}^{\beta}}{(1-a_X^2)(1-a_X e^{i\lambda})} \tag{A.32}$$

with $\alpha < \beta$ for $|a_X| < 1$. Then the TE is

$$I_{X\to Y}^{\mathrm{TE}} = \lim_{\tau\to\infty} \frac{1}{2} \ln \frac{1}{\sigma_Y^2} \frac{|G_\tau|}{|G_{\tau-1}|}$$

$$= \lim_{\tau\to\infty} \frac{1}{2} \ln \frac{|G_\tau|}{|G_{\tau-1}|} - \frac{1}{2}\ln\sigma_Y^2 \tag{A.33}$$

$$= \frac{1}{2} \ln \lim_{\tau\to\infty} \frac{|G_\tau|}{|G_{\tau-1}|} - \frac{1}{2}\ln\sigma_Y^2 \tag{A.34}$$

$$= \frac{1}{2} \ln \exp\left(\frac{1}{2\pi}\int_0^{2\pi} \ln f(\lambda) d\lambda\right) - \frac{1}{2}\ln\sigma_Y^2 \tag{A.35}$$

$$= \frac{1}{4\pi}\int_0^{2\pi} \ln f(\lambda) d\lambda - \frac{1}{2}\ln \sigma_Y^2 \tag{A.36}$$

$$= \frac{1}{4\pi}\Bigg[\underbrace{\int_0^{2\pi} \ln\left(\alpha e^{i\lambda} + \beta\right) d\lambda}_{(\star)} - \ln(1-a_X^2)\underbrace{\int_0^{2\pi} d\lambda}_{2\pi} - \underbrace{\int_0^{2\pi} \ln\left(1 - a_X e^{i\lambda}\right) d\lambda}_{(\star\star)}\Bigg] - \frac{1}{2}\ln\sigma_Y^2, \tag{A.37}$$

where the integrals $(\star)$ and $(\star\star)$ can be evaluated using contour integration to

$$(\star) = 2\pi \ln\beta = 2\pi \ln\left(c_{XY}^2\sigma_X^2 + \sigma_Y^2(1-a_X^2)\right) \qquad \text{for } \alpha \leq \beta, \tag{A.38}$$

$$(\star\star) = 2\pi \ln 1 = 0 \qquad \text{for } a_X \leq 1. \tag{A.39}$$

The TE is thus

$$I_{X\to Y}^{\mathrm{TE}} = \frac{1}{2}\ln\left(1 + \frac{(c_{XY}^2\sigma_X^2)/(1-a_X^2)}{\sigma_Y^2}\right) \tag{A.40}$$

and depends on the autodependency strength of X.

Case $a_X = 0$

Now the process "decouples in time" since no autodependencies are present. The covariance matrix is

$$\Gamma_{Y_t Y_t^- W_t^- Z_t^-} = \begin{pmatrix} \Gamma_{Y_t} & 0 & \Gamma_{Y_t;W_t^-} & \Gamma_{Y_t;Z_t^-} \\ 0 & \Gamma_{Y_t^-} & \Gamma_{Y_t^-;W_t^-} & \Gamma_{Y_t^-;Z_t^-} \\ \Gamma_{Y_t;W_t^-}^\top & \Gamma_{Y_t^-;W_t^-}^\top & \Gamma_{W_t^-} & 0 \\ \Gamma_{Y_t;Z_t^-}^\top & \Gamma_{Y_t^-;Z_t^-}^\top & 0 & \Gamma_{Z_t^-} \end{pmatrix}, \tag{A.41}$$

with the blocks being

$$\begin{aligned} \Gamma_{Y_t} &= c_{WY}^2\sigma_W^2 + c_{XY}^2\sigma_X^2 + \sigma_Y^2 \\ \Gamma_{Y_t;W_t^-} &= (c_{WY}\sigma_W^2, 0, 0, \ldots) \\ \Gamma_{Y_t;Z_t^-} &= (c_{XY}c_{XZ}\sigma_X^2, 0, 0, \ldots) \\ \Gamma_{Y_t^-} &= (c_{WY}^2\sigma_W^2 + c_{XY}^2\sigma_X^2 + \sigma_Y^2)\mathbb{I} \\ \Gamma_{Y_t^-;W_t^-} &= c_{WY}\sigma_W^2\mathbb{S} \\ \Gamma_{Y_t^-;Z_t^-} &= c_{XY}c_{XZ}\sigma_X^2\mathbb{S} \end{aligned}$$

$$\Gamma_{W_t^-} = \sigma_W^2 \mathbb{I}$$
$$\Gamma_{Z_t^-} = (c_{XZ}^2 \sigma_X^2 + \sigma_Z^2)\mathbb{I}\,,$$

where $\mathbb{I}$ is the identity matrix and $\mathbb{S}$ is the shift matrix with ones on the superdiagonal, i.e., the first upper off-diagonal, and zeros everywhere else. The quotient in Eq. (A.22) can be simplified by expressing the block matrix in terms of the Schur complement of the covariance block $\Gamma_{Y_t^- W_t^- Z_t^-}$

$$\frac{|\Gamma_{Y_t Y_t^- W_t^- Z_t^-}|}{|\Gamma_{Y_t^- W_t^- Z_t^-}|} = \left| \Gamma_{Y_t} - (\Gamma_{Y_t;Y_t^-}, \Gamma_{Y_t;W_t^-}, \Gamma_{Y_t;Z_t^-})(\Gamma_{Y_t^- W_t^- Z_t^-})^{-1} \begin{pmatrix} \Gamma_{Y_t;Y_t^-}^\top \\ \Gamma_{Y_t;W_t^-}^\top \\ \Gamma_{Y_t;Z_t^-}^\top \end{pmatrix} \right| . \quad \text{(A.42)}$$

Since the vector $(\Gamma_{Y_t;Y_t^-}, \Gamma_{Y_t;W_t^-}, \Gamma_{Y_t;Z_t^-})$ contains only two non-zero elements, we do not have to take the infinite limit and do not need to invert the whole matrix $\Gamma_{Y_t^- W_t^- Z_t^-}$. A simple calculation yields

$$\frac{|\Gamma_{Y_t Y_t^- W_t^- Z_t^-}|}{|\Gamma_{Y_t^- W_t^- Z_t^-}|} = c_{WY}^2 \sigma_W^2 + c_{XY}^2 \sigma_X^2 + \sigma_Y^2 - \frac{c_{WY}^2 \sigma_W^4}{\sigma_W^2} - \frac{c_{XY}^2 c_{XZ}^2 \sigma_X^4}{c_{XZ}^2 \sigma_X^2 + \sigma_Z^2}, \quad \text{(A.43)}$$

from which we get

$$I_{X \to Y}^{\mathrm{TE}} = \frac{1}{2} \ln \left(1 + \frac{c_{XY}^2 \sigma_X^2 \sigma_Z^2}{\sigma_Y^2 (c_{XZ}^2 \sigma_X^2 + \sigma_Z^2)} \right). \quad \text{(A.44)}$$

Here, the TE depends on the coupling strength of X with Z, which seems rather unintuitive. This formula could have also been derived by first exploiting separation properties of the corresponding time series graph (i.e., the Markov property of the process), from which a much smaller set of conditions can be inferred.

A.3.2. Derivation for MIT

The measures based on the parental sets are much easier to derive because they involve only finite and very low dimensional covariance matrices. As an example, for the entropy $H(Y_t \mid W_{t-1}, X_{t-3})$ needed to compute the MIT, the covariance matrix of (Y_t, W_{t-1}, X_{t-3}) is

$$\begin{pmatrix} c_{WY}^2 \sigma_W^2 + \frac{c_{XY}^2 \sigma_X^2}{1-a_X^2} + \sigma_Y^2 & c_{WY} \sigma_W^2 & \frac{a_X c_{XY} \sigma_X^2}{1-a_X^2} \\ c_{WY} \sigma_W^2 & \sigma_W^2 & 0 \\ \frac{a_X c_{XY} \sigma_X^2}{1-a_X^2} & 0 & \frac{\sigma_X^2}{1-a_X^2} \end{pmatrix}. \quad \text{(A.45)}$$

A.4. Proofs

A.4.1. Proof of inequality theorem

The MIT $I^{\text{MIT}}_{X\to Y} = I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}})$ between two uni- or multivariate subcomponents X, Y of a stationary multivariate discrete-time stochastic process $\mathbf{X}$ with time series graph G and parents $\mathcal{P}$ is bounded by the two CMIs with condition on either parents [Eq. (5.38)]

$$I(X_{t-\tau}; Y_t|\mathcal{P}_{X_{t-\tau}}) \;\leq\; I^{\text{MIT}}_{X\to Y} \;\leq\; I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}). \tag{A.46}$$

where $\tau > 0$. The right inequality holds for all Markov processes and the left inequality if additionally the "no sidepath"-condition (5.37) for the coupling "$X_{t-\tau} \to Y_t$" holds, that is, if $X_{t-\tau}$ is separated from $\mathcal{P}_{X_{t-\tau}} \setminus \mathcal{P}_{Y_t}$ by its parents $\mathcal{P}_{X_{t-\tau}}$ in the time series graph.

Proof. To prove the right inequality, let $\tilde{\mathcal{P}}_{X_{t-\tau}}$ be the set of parents of $X_{t-\tau}$ that is not already included in $\mathcal{P}_{Y_t}$, i.e., $\tilde{\mathcal{P}}_{X_{t-\tau}} = \mathcal{P}_{X_{t-\tau}} \setminus \mathcal{P}_{Y_t}$. Then it holds that $I(\tilde{\mathcal{P}}_{X_{t-\tau}}; Y_t|\mathcal{P}_{Y_t}) = 0$ because the parents $\mathcal{P}_{Y_t}$ separate Y_t from any subset of $\mathbf{X}^-_t \setminus \mathcal{P}_{Y_t}$. Now we apply the chain rule on the (multivariate) CMI $I(X_{t-\tau}, \tilde{\mathcal{P}}_{X_{t-\tau}}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\})$ twice:

$$\begin{aligned}
&I(X_{t-\tau}, \tilde{\mathcal{P}}_{X_{t-\tau}}; Y_t|\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\}) = \\
&= I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\}) + \underbrace{I(\tilde{\mathcal{P}}_{X_{t-\tau}}; Y_t|\mathcal{P}_{Y_t})}_{=0} && \text{(A.47)}\\
&= \underbrace{I(\tilde{\mathcal{P}}_{X_{t-\tau}}; Y_t|\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\})}_{\geq 0} + I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\}, \tilde{\mathcal{P}}_{X_{t-\tau}}) && \text{(A.48)}\\
&\implies I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}) \leq I(X_{t-\tau}; Y_t|\mathcal{P}_{Y_t}\backslash\{X_{t-\tau}\}). && \text{(A.49)}
\end{aligned}$$

For the left inequality we now define $\tilde{\mathcal{P}}_{Y_t}$ to be the set of parents of Y_t that is not already included in $\mathcal{P}_{X_{t-\tau}}$, i.e., $\tilde{\mathcal{P}}_{Y_t} = \mathcal{P}_{Y_t} \setminus \mathcal{P}_{X_{t-\tau}}$. Then under the "no sidepath"-condition it holds that $I(\tilde{\mathcal{P}}_{Y_t} \setminus \{X_{t-\tau}\}; X_{t-\tau}|\mathcal{P}_{X_{t-\tau}}) = 0$. Note that all paths emanating from $X_{t-\tau}$ towards the past are surely blocked by $\mathcal{P}_{X_{t-\tau}}$ because they contain the motifs "$\to Z_{t-\tau'} \to X_{t-\tau}$" or "$-Z_{t-\tau'} \to X_{t-\tau}$" which are both blocked as $Z_{t-\tau'} \in \mathcal{P}_{X_{t-\tau}}$. The "no sidepath"-condition further demands that there are no unblocked paths to $\tilde{\mathcal{P}}_{Y_t}$ emanating towards the present or future. Again, we apply the chain rule on the (multivariate) CMI $I(X_{t-\tau}; Y_t, \tilde{\mathcal{P}}_{Y_t} \setminus \{X_{t-\tau}\}|\mathcal{P}_{X_{t-\tau}})$ twice:

$$\begin{aligned}
&I(X_{t-\tau}; Y_t, \tilde{\mathcal{P}}_{Y_t} \setminus \{X_{t-\tau}\}|\mathcal{P}_{X_{t-\tau}}) = \\
&= I(X_{t-\tau}; Y_t|\mathcal{P}_{X_{t-\tau}}) + \underbrace{I(X_{t-\tau}; \tilde{\mathcal{P}}_{Y_t} \setminus \{X_{t-\tau}\}|\mathcal{P}_{X_{t-\tau}}, Y_t)}_{\geq 0} && \text{(A.50)}\\
&= \underbrace{I(\tilde{\mathcal{P}}_{Y_t} \setminus \{X_{t-\tau}\}; X_{t-\tau}|\mathcal{P}_{X_{t-\tau}})}_{=0} + I(X_{t-\tau}; Y_t|\tilde{\mathcal{P}}_{Y_t} \setminus \{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}) && \text{(A.51)}
\end{aligned}$$

$$\implies I(X_{t-\tau}; Y_t | \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}) \geq I(X_{t-\tau}; Y_t | \mathcal{P}_{X_{t-\tau}}). \tag{A.52}$$

□

A.4.2. Proof of coupling strength autonomy theorems

Momentary information transfer

The proof essentially utilizes the data processing inequality of CMI (Eq. (3.23)), translational invariance (Eq. (3.26)), and the Markov property (Eq. (3.27)).

Proof. Combining the expressions for X and Y in the additivity condition (5.35) and inserting them into the definition of MIT, we get

$$\begin{aligned}
I^{\text{MIT}}_{X \to Y}(\tau) &\equiv I(X_{t-\tau}; Y_t | \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\}, \mathcal{P}_{X_{t-\tau}}) && \text{(A.53)} \\
&\overset{\text{Eq. (5.35)}}{=} I(g_X(\mathcal{P}_{X_{t-\tau}}) + \eta^X_{t-\tau};\ f(X_{t-\tau}) + g_Y(\mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\}) + \\
&\qquad + \eta^Y_t \mid \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\},\ \mathcal{P}_{X_{t-\tau}}) && \text{(A.54)} \\
&\overset{\text{Eq. (3.26)}}{=} I(\eta^X_{t-\tau};\ f(X_{t-\tau}) + \eta^Y_t \mid \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\},\ \mathcal{P}_{X_{t-\tau}}), && \text{(A.55)}
\end{aligned}$$

giving Eq. (5.41). If we further assume linearity of f, condition (5.36), this can be simplified to

$$\begin{aligned}
I^{\text{MIT}}_{X \to Y}(\tau) &\overset{\text{Eq. (5.36)}}{=} I(\eta^X_{t-\tau};\ c(\eta^X_{t-\tau} + g_X(\mathcal{P}_{X_{t-\tau}})) + \eta^Y_t \mid \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\},\ \mathcal{P}_{X_{t-\tau}}) && \text{(A.56)} \\
&\overset{\text{Eq. (3.26)}}{=} I(\eta^X_{t-\tau};\ c\eta^X_{t-\tau} + \eta^Y_t \mid \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\},\ \mathcal{P}_{X_{t-\tau}}) && \text{(A.57)} \\
&\overset{\text{Eq. (3.27)}}{=} I(\eta^X_{t-\tau};\ c\eta^X_{t-\tau} + \eta^Y_t \mid \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\}), && \text{(A.58)}
\end{aligned}$$

arriving at Eq. (5.40). Finally, the "no sidepath"-condition (5.36) implies that $\eta^X_{t-\tau}$ is independent of $\mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\}$. Since also η^Y_t is independent of any past process, we can drop the condition due to the Markov property,

$$I^{\text{MIT}}_{X \to Y}(\tau) \overset{\text{Eq. (3.27)}}{=} I(\eta^X_{t-\tau};\ c\eta^X_{t-\tau} + \eta^Y_t), \tag{A.59}$$

arriving at Eq. (5.39).

For the contemporaneous MIT, the additivity condition (5.35) yields

$$\begin{aligned}
I^{\text{MIT}}_{X - Y} &\equiv I(X_t; Y_t | \mathcal{P}_{Y_t}, \mathcal{P}_{X_t}, \mathcal{N}_{X_t} \backslash \{Y_t\}, \mathcal{N}_{Y_t} \backslash \{X_t\}, \mathcal{P}(\mathcal{N}_{X_t} \backslash \{Y_t\}), \mathcal{P}(\mathcal{N}_{Y_t} \backslash \{X_t\})) && \text{(A.60)} \\
&\overset{\text{Eq. (5.35)}}{=} I(\eta^X_t + g_X(\mathcal{P}_{X_t}); \eta^Y_t + g_Y(\mathcal{P}_{Y_t}) | \mathcal{P}_{Y_t}, \mathcal{P}_{X_t}, \mathcal{N}_{X_t} \backslash \{Y_t\}, \mathcal{N}_{Y_t} \backslash \{X_t\}, \\
&\qquad \mathcal{P}(\mathcal{N}_{X_t} \backslash \{Y_t\}), \mathcal{P}(\mathcal{N}_{Y_t} \backslash \{X_t\})) && \text{(A.61)} \\
&\overset{\text{Eq. (3.26)}}{=} I(\eta^X_t; \eta^Y_t | \mathcal{P}_{Y_t}, \mathcal{P}_{X_t}, \mathcal{N}_{X_t} \backslash \{Y_t\}, \mathcal{N}_{Y_t} \backslash \{X_t\}, \mathcal{P}(\mathcal{N}_{X_t} \backslash \{Y_t\}), \mathcal{P}(\mathcal{N}_{Y_t} \backslash \{X_t\})) && \text{(A.62)}
\end{aligned}$$

$$\overset{\text{Eq. (3.27)}}{=} I(\eta_t^X; \eta_t^Y | \mathcal{N}_{X_t} \backslash \{Y_t\}, \mathcal{N}_{Y_t} \backslash \{X_t\}) \,, \tag{A.63}$$

arriving at Eq. (5.42). □

Momentary information transfer along paths

In the theorem, we denoted those parents of Y that are in the path nodes $\mathcal{C}_{X \to Y}$ defined in Eq. (2.13) as $\mathcal{P}_Y^{\mathcal{C}} = \mathcal{P}_Y \cap \mathcal{C}_{X \to Y}$ and correspondingly for other path nodes. Also note that $X_{t-\tau}$ is included in the set of path nodes.

Proof. We insert the dependencies assumed for X and Y in Eq. (5.45) in the definition of MITP (Eq. (3.53)) with the time indices dropped:

$$I^{\text{MITP}}_{X \to Y} \equiv I(X; Y \mid \mathcal{P}_Y \backslash \mathcal{C}_{X,Y}, \mathcal{P}(\mathcal{C}_{X \to Y})) \tag{A.64}$$

$$\overset{\text{Eq. (5.45)}}{=} I(g_X(\mathcal{P}_X) + \eta^X; f_Y(\mathcal{P}_Y^{\mathcal{C}}) + g_Y(\mathcal{P}_Y \setminus \mathcal{P}_Y^{\mathcal{C}}) + \eta^Y \mid \mathcal{P}_Y \backslash \mathcal{C}_{X,Y}, \mathcal{P}(\mathcal{C}_{X \to Y})) \tag{A.65}$$

$$\overset{\text{Eq. (3.26)}}{=} I(\eta^X; f_Y(\mathcal{P}_Y^{\mathcal{C}}) + \eta^Y \mid \mathcal{P}_Y \backslash \mathcal{C}_{X,Y}, \mathcal{P}(\mathcal{C}_{X \to Y})). \tag{A.66}$$

In the theorem, f_Y is assumed linear and we also assumed all other path nodes $Z^{(i)} \in \mathcal{C}_{X \to Y}$ to linearly depend on each other by Eq. (5.46), where dependencies on external nodes were only assumed additive. Then,

$$I^{\text{MITP}}_{X \to Y} \overset{\text{Eq. (3.26)}}{=} I(\eta^X; f(\eta^X, \cup_i \eta^i) + \eta^Y \mid \mathcal{P}_Y \backslash \mathcal{C}_{X,Y}, \mathcal{P}(\mathcal{C}_{X \to Y})), \tag{A.67}$$

for some linear function f. Finally, since $\mathcal{C}_{X \to Y}$ entails nodes on all directed causal paths emanating from $X_{t-\tau}$ and ending in Y_t, also all sidepath nodes are included in this set. This implies that the innovations $\eta^X, \cup_i \eta^i$ are independent of the external parents, which can be expressed with a multivariate MI as

$$I\left((\eta^X, \eta^Y, \cup_i \eta^i) \; ; \; (\mathcal{P}_Y \backslash \mathcal{C}_{X,Y}, \mathcal{P}(\mathcal{C}_{X \to Y}))\right) = 0 \,, \tag{A.68}$$

and we can drop the condition due to the Markov property,

$$I^{\text{MITP}}_{X \to Y} \overset{\text{Eq. (3.27)}}{=} I(\eta^X; f(\eta^X, \cup_i \eta^i) + \eta^Y), \tag{A.69}$$

yielding Eq. (5.47). □

Momentary interaction information

Using the same assumptions as for Theorem 5.4, the dependencies of momentary interaction information between $X_{t-\tau}$, Y_t and an intermediate process $Z_{t-\tau_Z} \in \mathcal{C}_{X \to Y} \setminus \{X_{t-\tau}\}$ can be simplified exploiting the same arguments as above.

Proof.

$$\mathcal{I}^{\text{MII}}_{X\to Y|Z}(\tau,\tau_Z) \equiv \mathcal{I}(X_{t-\tau};Y_t;Z_{t-\tau_Z} \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t},\mathcal{P}(\mathcal{C}_{X_{t-\tau}\to Y_t})) \tag{A.70}$$

$$\overset{\text{Eq. (3.26)}}{\equiv} \mathcal{I}(\eta^X; f(\eta^X,\cup_i\eta^i)+\eta^Y; f_Z(\eta^X,\cup_i\eta^i\setminus\eta^Z)+\eta^Z \mid \mid \mathcal{P}_{Y_t}\backslash\mathcal{C}_{X_{t-\tau},Y_t},\mathcal{P}(\mathcal{C}_{X_{t-\tau}\to Y_t})) \tag{A.71}$$

$$\overset{\text{Eq. (3.27)}}{\equiv} \mathcal{I}(\eta^X; f(\eta^X,\cup_i\eta^i)+\eta^Y; f_Z(\eta^X,\cup_i\eta^i\setminus\eta^Z)+\eta^Z), \tag{A.72}$$

giving Eq. (5.48) with linear functions f, f_Z. □

A.5. Interactions between three processes – all cases

In Sect. 5.2.4, we discussed momentary interaction information only for an example of three process, which are coupled solely by directed links shown in Fig. 5.5. Here, we discuss *all* four possible cases. In Fig. A.1 we show the example time series graphs that include all possible combinations of causal and contemporaneous links between three processes. Essentially, they represent the four unconditioned open motifs defined in Fig. 2.5. These results yield some insights on the information-theoretic aspects of Gaussian channels between multiple processes.

First, we need to define two more versions of MII for two different cases of contemporaneous links between the three processes $X_{t-\tau}$, $Z_{t-\tau_Z}$, and Y_t. In Sect. 3.5.2, we defined MII for $\tau > \tau_Z > 0$, where the process $Z_{t-\tau_Z}$ acts as an intermediate process on a causal path. For the case $\tau = 0$ and $\tau_Z > 0$, Z acts as a common driver of the contemporaneous interaction between X and Y and for $\tau = \tau_Z = 0$, MII measures the contemporaneous momentary interaction. In analogy to MIT for $\tau = 0$ and/or $\tau_Z = 0$, the contemporaneous processes are additionally conditioned on the corresponding neighbors and parents of neighbors:

$$\begin{aligned}&\mathcal{I}^{\text{MII}}_{X-Y|Z}(\tau=0,\tau_z>0) \equiv \\ &\mathcal{I}(X_t;Y_t;Z_{t-\tau_Z}|\mathcal{P}_{Y_t}\backslash\{Z_{t-\tau_Z}\},\mathcal{P}_{X_t}\backslash\{Z_{t-\tau_Z}\},\mathcal{P}_{Z_{t-\tau_Z}},\mathcal{N}_{X_t}\backslash\{Y_t\},\mathcal{N}_{Y_t}\backslash\{X_t\}, \\ &\qquad \{\mathcal{P}(\mathcal{N}_{X_t}\backslash\{Y_t\}),\mathcal{P}(\mathcal{N}_{Y_t}\backslash\{X_t\}\}\backslash\{Z_{t-\tau_Z}\}),\end{aligned} \tag{A.73}$$

$$\begin{aligned}&\mathcal{I}^{\text{MII}}_{X-Y|Z}(\tau=0,\tau_z=0) \equiv \\ &\mathcal{I}(X_t;Y_t;Z_t|\mathcal{P}_{Y_t},\mathcal{P}_{X_t},\mathcal{P}_{Z_t},\mathcal{N}_{X_t}\backslash\{Y_t,Z_t\},\mathcal{N}_{Y_t}\backslash\{X_t,Z_t\},\mathcal{N}_{Z_t}\backslash\{X_t,Y_t\} \\ &\qquad \mathcal{P}(\mathcal{N}_{X_t}\backslash\{Y_t,Z_t\},\mathcal{N}_{Y_t}\backslash\{X_t,Z_t\},\mathcal{N}_{Z_t}\backslash\{X_t,Y_t\})),\ .\end{aligned} \tag{A.74}$$

Now, we consider the four cases drawn in Fig. A.1. To simplify notation, we drop the time indices and assume no further sidepaths, which can be expressed with a multivariate MI as

$$I\left((\eta^X,\eta^Y,\eta^Z)\ ;\ (\mathcal{P}_Y\setminus\{X,Z\},\mathcal{P}_X\setminus\{Z\})\right)=0. \tag{A.75}$$

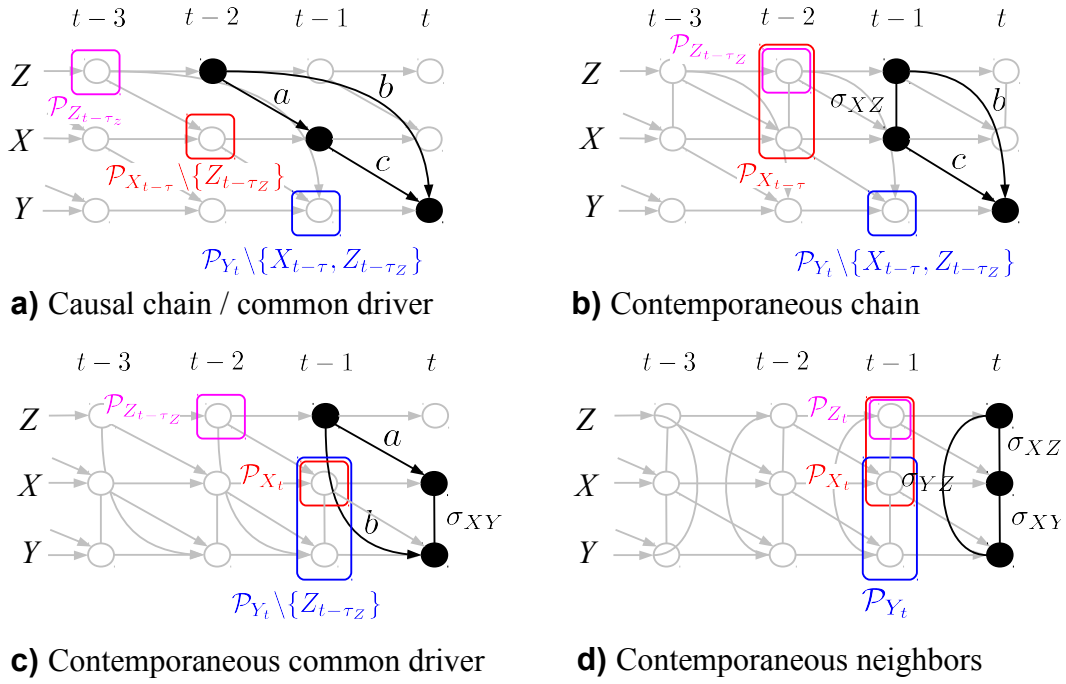


Figure A.1.: Example time series graphs that include all possible combinations of causal and contemporaneous links between three processes (black dots). The parents of Y_t, $X_{t-\tau}$ and $Z_{t-\tau_Z}$ are shown as blue, red and magenta boxes, respectively. a, b, c, σ_{XY}, σ_{XZ}, σ_{YZ} denote the coefficients for the Gaussian examples analyzed in the text.

A.5.1. Directed causal chain / common driver dependency

For completeness, we recapitulate also the first case. Consider the momentary interaction information (MII, Eq. (3.55)) of the three black nodes in Fig. A.1(a) which can be expressed using the additivity and linearity assumptions as

$$\begin{aligned} Z &= g_Z(\mathcal{P}_Z) + \eta^Z \\ X &= aZ + g_X(\mathcal{P}_X \setminus \{Z\}) + \eta^X \\ Y &= cX + bZ + g_Y(\mathcal{P}_Y \setminus \{X, Z\}) + \eta^Y \,, \end{aligned} \tag{A.76}$$

with η^Z, η^Y, η^Z being independent. Then, with the same arguments as in Sect. A.4.2, MII given by Eq. (3.55) reduces to

$$\begin{aligned} \mathcal{I}^{\text{MII}}_{X\to Y|Z} &= \mathcal{I}(\eta^Y + c(\eta^X + a\eta^Z) + b\eta^Z; \eta^X + a\eta^Z; \eta^Z) && \text{(A.77)} \\ &= I(\eta^Y + c(\eta^X + a\eta^Z) + b\eta^Z; \eta^X + a\eta^Z) - I(\eta^Y + c\eta^X; \eta^X). && \text{(A.78)} \end{aligned}$$

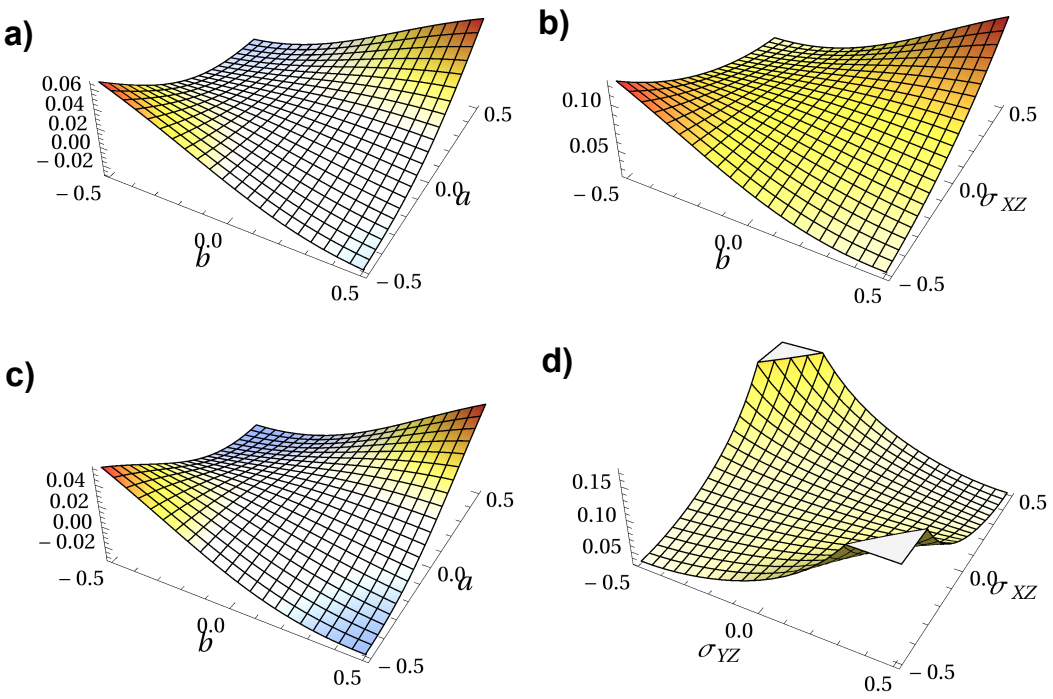


Figure A.2.: Momentary interaction information (MII) for the four cases shown in Fig. A.1 for Gaussian examples given by (a) Eq. (A.79) for varying a, b with fixed $c = 0.25$, (b) Eq. (A.80) for varying σ_{XZ}, b with fixed $c = 0.25$, (c) Eq. (A.84) for varying a, b with fixed $\sigma_{XY} = 0.25$, and (d) Eq. (A.88) for varying σ_{XZ}, σ_{YZ} with fixed $\sigma_{XY} = 0.25$. All innovation terms have unit variances. The shading denotes the value of the vertical axis. Blue shadings correspond to negative interaction information.

The latter term being the MIT from Eq. (5.39). If all processes are jointly Gaussian with coefficients denoted in the figure and innovation's variances σ_X^2, σ_Y^2, σ_Z^2, their MII at the causal lags τ=1, τ_Z=2 is given by

$$\mathcal{I}^{\text{MII}}_{X\to Y|Z} = \frac{1}{2}\ln\left(\frac{(\sigma_Y^2 + c^2\sigma_X^2 + (ac+b)^2\sigma_Z^2)(\sigma_X^2 + a^2\sigma_Z^2)}{((\sigma_Y^2 + b^2\sigma_Z^2)(\sigma_X^2 + a^2\sigma_Z^2) - a^2b^2\sigma_Z^4)} \times \frac{\sigma_Y^2}{(\sigma_Y^2 + c^2\sigma_X^2)}\right), \quad \text{(A.79)}$$

repeating Eq. (5.22). This case was discussed in Sect. 5.2.4.

A.5.2. Contemporaneous chain

Figure A.1(b) shows this case with $(\eta^Z, \eta^X) \perp\!\!\!\perp \eta^Y$, but η^Z, η^X possibly dependent:

$$Z = g_Z(\mathcal{P}_Z) + \eta^Z$$

$$
\begin{aligned}
X &= g_X(\mathcal{P}_X) + \eta^X \\
Y &= cX + bZ + g_Y(\mathcal{P}_Y \setminus \{X, Z\}) + \eta^Y .
\end{aligned} \tag{A.80}
$$

Here, MII given by Eq. (3.55) reduces to

$$
\mathcal{I}^{\mathrm{MII}}_{X \to Y|Z} = \mathcal{I}(\eta^X; c\eta^X + b\eta^Z + \eta^Y; \eta^Z) \tag{A.81}
$$

$$
= I(\eta^X; c\eta^X + b\eta^Z + \eta^Y) - I(\eta^X; c\eta^X + \eta^Y | \eta^Z). \tag{A.82}
$$

The latter term can only be simplified if η^Z, η^X are independent. In the Gaussian case, MII is given by

$$
\begin{aligned}
\mathcal{I}^{\mathrm{MII}}_{X \to Y|Z} = \frac{1}{2} \ln \Bigg(& \frac{\sigma_X^2 \left(b^2\sigma_Z^2 + bc\sigma_{XZ} + c^2\sigma_X^2 + \sigma_Y^2\right)}{\left(\sigma_X^2 \left(b^2\sigma_Z^2 + bc\sigma_{XZ} + c^2\sigma_X^2 + \sigma_Y^2\right) - \left(b\sigma_{XZ} + c\sigma_X^2\right)^2\right)} \times \\
& \times \frac{\left(\sigma_X^2\sigma_Z^2 - \sigma_{XZ}^2\right)\left(\sigma_Z^2\left(c^2\sigma_X^2 + \sigma_Y^2\right) - c^2\sigma_{XZ}^2\right)}{\sigma_Z^2\left(\sigma_X^2\sigma_Y^2\sigma_Z^2 - \sigma_{XZ}^2\sigma_Y^2\right)} \Bigg).
\end{aligned} \tag{A.83}
$$

Here the dependence of MII on the coefficients a, b is similar to the previous case with the difference that MII cannot become negative (can be seen using the triangle inequality $\sigma_X^2\sigma_Y^2 \geq \sigma_{XY}^2$), that is, the link $X \to Y$ cannot be counteracted by Z. Intuitively, since Z is not causally driving X or can change the interaction between X and Y as an intermediate node, it cannot counteract.

A.5.3. Contemporaneous driver

Figure A.1(c) shows a case where Z acts as a common driver on a contemporaneous link with $(\eta^X, \eta^Y) \perp\!\!\!\perp \eta^Z$, but η^X, η^Y possibly dependent:

$$
\begin{aligned}
Z &= g_Z(\mathcal{P}_Z) + \eta^Z \\
X &= aZ + g_X(\mathcal{P}_X \setminus \{Z\}) + \eta^X \\
Y &= bZ + g_Y(\mathcal{P}_Y \setminus \{X, Z\}) + \eta^Y .
\end{aligned} \tag{A.84}
$$

Then MII, given by Eq. (A.73), reduces to

$$
\mathcal{I}^{\mathrm{MII}}_{X - Y|Z} = \mathcal{I}(a\eta^Z + \eta^X; b\eta^Z + \eta^Y; \eta^Z \mid \mathcal{N}_Y \backslash \{X\}, \mathcal{N}_X \backslash \{Y\}) \tag{A.85}
$$

$$
\begin{aligned}
&= I(a\eta^Z + \eta^X; b\eta^Z + \eta^Y \mid \mathcal{N}_Y \backslash \{X\}, \mathcal{N}_X \backslash \{Y\}) \\
&\quad - I(\eta^X; \eta^Y \mid \mathcal{N}_Y \backslash \{X\}, \mathcal{N}_X \backslash \{Y\}).
\end{aligned} \tag{A.86}
$$

The latter term being the MIT from Eq. (5.42). This case is similar to the first case and for multivariate Gaussians, MII is given by

$$\mathcal{I}^{\text{MII}}_{X-Y|Z} = \frac{1}{2}\ln\left(\frac{(a^2\sigma_Z^2+\sigma_X^2)(b^2\sigma_Z^2+\sigma_Y^2)}{(a^2\sigma_Z^2+\sigma_X^2)(b^2\sigma_Z^2+\sigma_Y^2)-\left(ab\sigma_Z^2+\sigma_{XY}\right)^2}\times\frac{\left(\sigma_X^2\sigma_Y^2-\sigma_{XY}^2\right)}{\sigma_X^2\sigma_Y^2}\right), \tag{A.87}$$

and can also be negative if a and b are of different sign.

A.5.4. Contemporaneous neighbors

Last, Fig. A.1(d) shows the non-causal case of three contemporaneous neighbors with η^X, η^Y, η^Z possibly dependent:

$$\begin{aligned} Z &= g_Z(\mathcal{P}_Z)+\eta^Z \\ X &= g_X(\mathcal{P}_X)+\eta^X \\ Y &= g_Y(\mathcal{P}_Y)+\eta^Y, \end{aligned} \tag{A.88}$$

for which MII, given by Eq. (A.74), reduces to

$$\mathcal{I}^{\text{MII}}_{X-Y|Z} = \mathcal{I}(\eta^X;\eta^Y;\eta^Z \mid \mathcal{N}_Y\backslash\{X,Z\},\mathcal{N}_X\backslash\{Y,Z\},\mathcal{N}_Z\backslash\{X,Y\}) \tag{A.89}$$

$$\begin{aligned} &= I(\eta^X;\eta^Y \mid \mathcal{N}_Y\backslash\{X,Z\},\mathcal{N}_X\backslash\{Y,Z\},\mathcal{N}_Z\backslash\{X,Y\}) \\ &\quad - I(\eta^X;\eta^Y \mid \mathcal{N}_Y\backslash\{X\},\mathcal{N}_X\backslash\{Y\},\mathcal{N}_Z\backslash\{X,Y\}). \end{aligned} \tag{A.90}$$

In the Gaussian case,

$$\begin{aligned} \mathcal{I}^{\text{MII}}_{X-Y|Z} = \frac{1}{2}\ln\Bigg(&\frac{\sigma_X^2\sigma_Y^2}{\left(\sigma_X^2\sigma_Y^2-\sigma_{XY}^2\right)}\times \\ &\times\frac{\left(\sigma_X^2\sigma_Z^2-\sigma_{XZ}^2\right)\left(\sigma_Y^2\sigma_Z^2-\sigma_{YZ}^2\right)}{\sigma_Z^2\left(\sigma_X^2\sigma_Y^2\sigma_Z^2+2\sigma_{XY}\sigma_{YZ}\sigma_{XZ}-\sigma_{XZ}^2\sigma_Y^2-\sigma_{YZ}^2\sigma_X^2-\sigma_{XY}^2\sigma_Z^2\right)}\Bigg). \end{aligned} \tag{A.91}$$

This MII can also not become negative. Since no causal links are involved here, it should not be interpreted as an influence of either process on the others and is merely a certain description of contemporaneous dependencies. For multivariate Gaussians, it is the difference between interpreting the covariance matrix of the innovation terms or its inverse.

A.6. Further results for linear theory

Here, we provide some more results for the linear theory of coupling strength. We give an interpretation of covariance in terms of parents and state the coupling strength autonomy theorem for MIT in the linear case. While one could object, that the linear case follows from the general case, it is nevertheless instructive and allows to interpret

MIT using concepts from linear theory common in classical statistics. The results here are from Runge (2013).

A.6.1. Interpretation of covariance in terms of parents

Complementing the path-theoretical analysis of covariance in Sect. 5.2.6, one can also characterize the dependencies of the covariance Eq. (A.10) in terms of the parents in the time series graph. Two univariate subprocesses X, Y of $\mathbf{X}$ given by Eq. (2.14),

$$\mathbf{X}_t = \sum_{s=1}^{p} \Phi(s)\mathbf{X}_{t-s} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \tag{A.92}$$

with a link "$X_{t-\tau} \to Y_t$" and $\tau > 0$ can be written as

$$X_t = \sum_{i=1}^{N_X} a_i Z^i_{t-h_i} + \varepsilon_{X,t} \tag{A.93}$$

$$Y_t = cX_{t-\tau} + \sum_{i=1}^{N_Y} b_i W^i_{t-g_i} + \varepsilon_{Y,t} \tag{A.94}$$

with parents

$$Z^i_{t-h_i} \in \mathcal{P}_{X_t} \quad \text{for} \quad i = 1, \dots, N_X, \tag{A.95}$$

$$W^i_{t-g_i} \in \mathcal{P}_{Y_t} \backslash \{X_{t-\tau}\} \quad \text{for} \quad i = 1, \dots, N_Y. \tag{A.96}$$

Here the coefficient c corresponds to the entry $\Phi(\tau)_{YX}$.

To simplify notation, Eqns. (A.93, A.94) are expressed in vector notation

$$\begin{aligned} X_t &= \mathbf{Z}_t A + \varepsilon_{X,t} \\ Y_t &= cX_{t-\tau} + \mathbf{W}_t B + \varepsilon_{Y,t} \end{aligned} \tag{A.97}$$

where X_t, Y_t are scalar random processes, A and B are the coefficient vectors, and $\mathbf{Z}_t$, $\mathbf{W}_t$ are possibly multivariate random processes of dimension N_X and N_Y respectively,

$$\mathbf{Z}_t = (Z^1_{t-h_1}, \dots, Z^{N_X}_{t-h_{N_X}}), \tag{A.98}$$

$$\mathbf{W}_t = (W^1_{t-g_1}, \dots, Z^{N_Y}_{t-g_{N_Y}}). \tag{A.99}$$

In the following, t and τ will be dropped for ease of notation.

For the cross correlation between X and Y at lag τ, the covariance $E[Y^\top X]$ and the variances $E[Y^\top Y]$ and $E[X^\top X]$ are needed. While in the covariance expression Eq. (A.10) the dependencies are rather hidden, the vector notation allows to derive them simply by directly plugging in Eqns. (A.97) into the covariances and using only $E[\mathbf{W}^\top \varepsilon_Y] = E[\mathbf{Z}^\top \varepsilon_Y] = E[\mathbf{Z}^\top \varepsilon_X] = 0$ since ε_Y and ε_X are i.i.d. processes

independent from the past parents. Then the (co-)variances can be written in a compact way:

$$\begin{aligned} E[Y^\top X] &= c\sigma_X^2 \\ &\quad + cA^\top E[\mathbf{Z}^\top \mathbf{Z}]A \\ &\quad + B^\top E[\mathbf{W}^\top \mathbf{Z}]A \\ &\quad + B^\top E[\mathbf{W}^\top \varepsilon_X], \end{aligned} \tag{A.100}$$

$$\begin{aligned} E[Y^\top Y] &= \sigma_Y^2 + c^2\sigma_X^2 \\ &\quad + c^2 A^\top E[\mathbf{Z}^\top \mathbf{Z}]A + B^\top E[\mathbf{W}^\top \mathbf{W}]B \\ &\quad + c\left(B^\top E[\mathbf{W}^\top \mathbf{Z}]A + A^\top E[\mathbf{Z}^\top \mathbf{W}]B\right) \\ &\quad + c\left(B^\top E[\mathbf{W}^\top \varepsilon_X] + E[\varepsilon_X^\top \mathbf{W}]B\right), \end{aligned} \tag{A.101}$$

$$\begin{aligned} E[X^\top X] &= \sigma_X^2 \\ &\quad + A^\top E[\mathbf{Z}^\top \mathbf{Z}]A. \end{aligned} \tag{A.102}$$

One can see, that the covariance $E[Y^\top X]$ not only depends on the coefficient c, but also on the variance of the parents $\mathbf{Z}$ of X, the covariance among the parents of X and Y and the covariance of the innovation ε_X with the parents $\mathbf{W}$ of Y.

Also in this interpretation, we find that the value of the cross correlation cannot easily be related to the coefficient c of the link between X and Y in the time series graph and depends on the multiple interactions between the parents of X and Y in the multivariate process.

A.6.2. Linear coupling strength autonomy theorem (with regression lemma)

Regression lemma

First, we show that also a multivariate regression on the parents in the graph is easier to interpret. We saw in Tab. 5.1 that the regression recovers the coefficients of the model, without intermixing the coefficients as for the univariate regressions. In analogy to the coupling strength autonomy theorem, also for the regressions a similar theorem holds, in that also for a general multivariate autoregressive process given by Eq. (A.92), the regressors

$$\mathbf{U}_Y^{\mathrm{MIT}} \equiv \mathcal{P}_Y \tag{A.103}$$

can be shown to yield the corresponding coefficients in the lagged matrices $\Phi(s)$. Regression coefficients for the regression on $\mathbf{U}_Y^{\mathrm{MIT}}$ and the partial correlation measure MIT capture different aspects of a coupling mechanism. A regression coefficient of a parent $X_{t-\tau}$ gives the scale factor which determines the proportion of $X_{t-\tau}$ influencing Y_t. The partial correlation, on the other hand is a *normalized* measure

and can, thus, better be compared to the partial correlation of other processes with quite different innovation's variances.

As in Eq. (A.97), the equation for the subprocess Y can be written as

$$Y_t = \mathbf{W}_t B + \varepsilon_{Y,t}, \tag{A.104}$$

where X and the coefficient c occurring in Eq. (A.97) is collapsed into $\mathbf{W}$ and B, respectively.

Lemma A.1. *For the autoregressive model Eq. (A.104), a multivariate regression for the dependent variable Y on $\mathbf{U} = (\mathbf{W}, \mathbf{V})$, where $\mathbf{V}$ are other regressors that are not part of the parents, i.e., $\mathbf{V} \cap \mathbf{W} = \emptyset$ gives*

$$\begin{pmatrix} \mathbf{R}_W \\ \mathbf{R}_V \end{pmatrix} = \begin{pmatrix} \mathbf{B} \\ 0 \end{pmatrix}. \tag{A.105}$$

Proof. A regression on $\mathbf{U} = (\mathbf{W}, \mathbf{V})$ gives the coefficient vector

$$\begin{aligned} \begin{pmatrix} \mathbf{R}_W \\ \mathbf{R}_V \end{pmatrix} &\equiv (E[\mathbf{U}^\top \mathbf{U}])^{-1} E[\mathbf{U}^\top Y] \\ &= \begin{pmatrix} E[\mathbf{W}^\top \mathbf{W}] & E[\mathbf{W}^\top \mathbf{V}] \\ E[\mathbf{V}^\top \mathbf{W}] & E[\mathbf{V}^\top \mathbf{V}] \end{pmatrix}^{-1} \begin{pmatrix} E[\mathbf{W}^\top Y] \\ E[\mathbf{V}^\top Y] \end{pmatrix}. \end{aligned} \tag{A.106}$$

The inverse can be treated via the matrix inversion lemma

$$\begin{aligned} &\begin{pmatrix} E[\mathbf{W}^\top \mathbf{W}] & E[\mathbf{W}^\top \mathbf{V}] \\ E[\mathbf{V}^\top \mathbf{W}] & E[\mathbf{V}^\top \mathbf{V}] \end{pmatrix}^{-1} \\ &= \begin{pmatrix} S_V^{-1} & -(E[\mathbf{W}^\top \mathbf{W}])^{-1} E[\mathbf{W}^\top \mathbf{V}] S_W^{-1} \\ -(E[\mathbf{V}^\top \mathbf{V}])^{-1} E[\mathbf{V}^\top \mathbf{W}] S_V^{-1} & S_W^{-1} \end{pmatrix}, \end{aligned} \tag{A.107}$$

where $S_\cdot$ denotes the Schur complements

$$S_V = E[\mathbf{W}^\top \mathbf{W}] - E[\mathbf{W}^\top \mathbf{V}](E[\mathbf{V}^\top \mathbf{V}])^{-1} E[\mathbf{V}^\top \mathbf{W}] \tag{A.108}$$

$$S_W = E[\mathbf{V}^\top \mathbf{V}] - E[\mathbf{V}^\top \mathbf{W}](E[\mathbf{W}^\top \mathbf{W}])^{-1} E[\mathbf{W}^\top \mathbf{V}]. \tag{A.109}$$

S_V can be interpreted as the conditional variance of $\mathbf{W}$ given $\mathbf{V}$. S_V^{-1} can be further transformed using the Woodbury matrix identity

$$\begin{aligned} S_V^{-1} = (E[\mathbf{W}^\top \mathbf{W}])^{-1} &- (E[\mathbf{W}^\top \mathbf{W}])^{-1} E[\mathbf{W}^\top \mathbf{V}] \times \\ &\times \underbrace{(-E[\mathbf{V}^\top \mathbf{V}] + E[\mathbf{V}^\top \mathbf{W}](E[\mathbf{W}^\top \mathbf{W}])^{-1} E[\mathbf{W}^\top \mathbf{V}])^{-1}}_{=(-S_W^{-1})} \times \\ &\times E[\mathbf{V}^\top \mathbf{W}](E[\mathbf{W}^\top \mathbf{W}])^{-1}. \end{aligned} \tag{A.110}$$

The covariance vector in Eq. (A.106) can be simplified by

$$\begin{pmatrix} E[\mathbf{W}^\top Y] \\ E[\mathbf{V}^\top Y] \end{pmatrix} = \begin{pmatrix} E[\mathbf{W}^\top \mathbf{W}]B + E[\mathbf{W}^\top \varepsilon_Y] \\ E[\mathbf{V}^\top \mathbf{W}]B + E[\mathbf{V}^\top \varepsilon_Y] \end{pmatrix} \tag{A.111}$$

$$= \begin{pmatrix} E[\mathbf{W}^\top \mathbf{W}]B \\ E[\mathbf{V}^\top \mathbf{W}]B \end{pmatrix}, \tag{A.112}$$

where $E[\mathbf{V}^\top \varepsilon_Y] = E[\mathbf{W}^\top \varepsilon_Y] = 0$ because ε_Y is independent of past processes. Then the regression coefficient $\mathbf{R}_W$ given by

$$\mathbf{R}_W = S_V^{-1} E[\mathbf{W}^\top \mathbf{W}]\mathbf{B} - (E[\mathbf{W}^\top \mathbf{W}])^{-1} E[\mathbf{W}^\top \mathbf{V}] S_W^{-1} E[\mathbf{V}^\top \mathbf{W}]\mathbf{B} \tag{A.113}$$

can be simplified by inserting Eq. (A.110) from which it follows that

$$S_V^{-1} E[\mathbf{W}^\top \mathbf{W}]\mathbf{B} = \mathbf{B} + (E[\mathbf{W}^\top \mathbf{W}])^{-1} E[\mathbf{W}^\top \mathbf{V}] S_W^{-1} E[\mathbf{V}^\top \mathbf{W}]\mathbf{B}, \tag{A.114}$$

and thus $\mathbf{R}_W = \mathbf{B}$ which proves the first part of the claim.

To prove the second part, now the analogue of Eq. (A.110) for S_W^{-1} is inserted into

$$\mathbf{R}_V = S_W^{-1} E[\mathbf{V}^\top \mathbf{W}]\mathbf{B} - (E[\mathbf{V}^\top \mathbf{V}])^{-1} E[\mathbf{V}^\top \mathbf{W}] S_V^{-1} E[\mathbf{W}^\top \mathbf{W}]\mathbf{B}, \tag{A.115}$$

from which using

$$\begin{aligned} S_W^{-1} E[\mathbf{V}^\top \mathbf{W}]\mathbf{B} &= (E[\mathbf{V}^\top \mathbf{V}])^{-1} E[\mathbf{V}^\top \mathbf{W}]\mathbf{B} \times \\ &\times (E[\mathbf{V}^\top \mathbf{V}]])^{-1} E[\mathbf{V}^\top \mathbf{W}] S_V^{-1} \times \\ &\times E[\mathbf{W}^\top \mathbf{V}](E[\mathbf{V}^\top \mathbf{V}])^{-1} E[\mathbf{V}^\top \mathbf{W}]\mathbf{B}, \end{aligned} \tag{A.116}$$

and

$$E[\mathbf{W}^\top \mathbf{V}](E[\mathbf{V}^\top \mathbf{V}])^{-1} E[\mathbf{V}^\top \mathbf{W}] = E[\mathbf{W}^\top \mathbf{W}] - S_V \tag{A.117}$$

one arrives at $\mathbf{R}_V = 0$. □

Linear coupling strength autonomy theorem

For the partial correlation MIT, the dependencies are slightly more complex.

Theorem A.1. *For the autoregressive model Eq. (2.14), written in vector notation as Eq. (A.97), the partial correlation $\rho^{\mathrm{MIT}}_{X\to Y}(\tau)$ given by Eq. (3.48) written in vector notation as in Eq. (3.35) with $\mathbf{U} = (\mathbf{W}, \mathbf{Z})$ is comprised of the covariances and variances*

$$\begin{aligned} E[Y_{\mathbf{U}}^\top X_{\mathbf{U}}] &= c\sigma_X^2 \\ &\quad - cE[\varepsilon_X^\top \mathbf{W}] S_Z^{-1} E[\mathbf{W}^\top \varepsilon_X] \\ E[Y_{\mathbf{U}}^\top Y_{\mathbf{U}}] &= \sigma_Y^2 + c^2\sigma_X^2 \end{aligned}$$

$$
\begin{aligned}
&\quad - c^2 E[\varepsilon_X^\top \mathbf{W}] S_Z^{-1} E[\mathbf{W}^\top \varepsilon_X] \\
E[X_{\mathbf{U}}^\top X_{\mathbf{U}}] &= \sigma_X^2 \\
&\quad - E[\varepsilon_X^\top \mathbf{W}] S_Z^{-1} E[\mathbf{W}^\top \varepsilon_X]. \qquad \text{(A.118)}
\end{aligned}
$$

where S_Z denotes the Schur complement

$$S_Z = E[\mathbf{W}^\top \mathbf{W}] - E[\mathbf{W}^\top \mathbf{Z}](E[\mathbf{Z}^\top \mathbf{Z}])^{-1} E[\mathbf{Z}^\top \mathbf{W}]\,, \qquad \text{(A.119)}$$

and the covariance $E[\varepsilon_X^\top \mathbf{W}]$ for each parent process $\mathbf{W}_i$ in terms of the coefficient path matrices Ψ and the innovation's covariance Σ it can be written as:

$$(E[\varepsilon_X^\top \mathbf{W}])_i = \sum_{r=1}^{N} \Psi_{W_i r}(\tau - g_i)\Sigma_{rX}. \qquad \text{(A.120)}$$

The proof is given below. The (co-)variances are comprised of two parts. The first one is simply the cross correlation between $\varepsilon_{X,t-\tau}$ and $\varepsilon_{Y,t} + c\,\varepsilon_{X,t-\tau}$. The second part is due to dependencies between $\varepsilon_{X,t-\tau}$ and the parents of Y_t and non-zero only under certain conditions.

More precisely, the Schur complement S_Z can be interpreted as the conditional variance of $\mathbf{W}$ given $\mathbf{Z}$. On the other hand, the covariance $(E[\varepsilon_X^\top \mathbf{W}])_i$ is the linear combination of all paths of length $\tau - g_i$ emanating from X_t or $\mathbf{X}_{r,t}$ with $\Sigma_{rX} \neq 0$ to $W^i_{t+\tau-g_i}$. It can be understood as a "sidepath"-covariance and is zero if there are no such paths. Then, for $E[\varepsilon_X^\top \mathbf{W}] = 0$, the ρ^{MIT} becomes

$$\rho^{\text{MIT}}_{X\to Y} = \frac{c\sigma_X}{\sqrt{\sigma_Y^2 + c^2\sigma_X^2}}. \qquad \text{(A.121)}$$

Thus, if there are no sidepaths, the partial correlation measure MIT of a link "$X_{t-\tau} \to Y_t$" solely depends on the coefficient matrix entry $\Phi_{YX}(\tau)$ and the innovation's variances σ_X^2 and σ_Y^2. The MIT of an autoregressive process is, therefore, much better interpretable than the cross correlation as analyzed in Sects. 5.2.6 and A.6.1 since its value is attributable to the interaction between $\mathbf{X}^j$ and $\mathbf{X}^i$ alone, i.e., the link "$\mathbf{X}^j_{t-\tau} \to \mathbf{X}^i_t$" in the time series graph of $\mathbf{X}$ as dicussed in Sect. 5.3.2.

Proof. First X and Y are regressed on $\mathbf{U} = (\mathbf{W}, \mathbf{Z})$ yielding the residuals

$$Y_{\mathbf{U}} \equiv Y - \mathbf{U}(E[\mathbf{U}^\top \mathbf{U}])^{-1} E[\mathbf{U}^\top Y] \qquad \text{(A.122)}$$

$$X_{\mathbf{U}} \equiv X - \mathbf{U}(E[\mathbf{U}^\top \mathbf{U}])^{-1} E[\mathbf{U}^\top X]. \qquad \text{(A.123)}$$

Then the covariance and variances are

$$E[Y_{\mathbf{U}}^\top X_{\mathbf{U}}] = E[Y^\top X] - E[Y^\top \mathbf{U}](E[\mathbf{U}^\top \mathbf{U}])^{-1} E[\mathbf{U}^\top X] \qquad \text{(A.124)}$$

$$E[Y_{\mathbf{U}}^\top Y_{\mathbf{U}}] = E[Y^\top Y] - E[Y^\top \mathbf{U}](E[\mathbf{U}^\top \mathbf{U}])^{-1} E[\mathbf{U}^\top Y] \qquad \text{(A.125)}$$

$$E[X_{\mathbf{U}}^\top X_{\mathbf{U}}] = E[X^\top X] - E[X^\top \mathbf{U}](E[\mathbf{U}^\top \mathbf{U}])^{-1} E[\mathbf{U}^\top X]. \qquad \text{(A.126)}$$

The covariance can be evaluated as follows. First, writing

$$X = \mathbf{U}\begin{pmatrix}0\\A\end{pmatrix} + \varepsilon_X \tag{A.127}$$
$$Y = \mathbf{U}\begin{pmatrix}\mathbf{B}\\c\mathbf{A}\end{pmatrix} + c\varepsilon_X + \varepsilon_Y \tag{A.128}$$

the covariance $E[Y^\top X]$ is expressed in terms of $\mathbf{U}$ as

$$\begin{aligned}&E[Y^\top X] = \\&\left(\mathbf{B}^\top, c\mathbf{A}^\top\right) E[\mathbf{U}^\top\mathbf{U}]\begin{pmatrix}0\\\mathbf{A}\end{pmatrix} + cE[\varepsilon_X^\top\mathbf{U}]\begin{pmatrix}0\\\mathbf{A}\end{pmatrix} + \underbrace{E[\varepsilon_Y^\top\mathbf{U}]}_{=0}\begin{pmatrix}0\\\mathbf{A}\end{pmatrix}\\&+ \left(\mathbf{B}^\top, c\mathbf{A}^\top\right) E[\mathbf{U}^\top\varepsilon_X] + c\underbrace{E[\varepsilon_X^\top\varepsilon_X]}_{\sigma_X^2} + \underbrace{E[\varepsilon_Y^\top\varepsilon_X]}_{=0},\end{aligned} \tag{A.129}$$

where $E[\varepsilon_Y^\top\mathbf{U}] = E[\varepsilon_Y^\top\varepsilon_X] = 0$ because ε_Y is i.i.d. and therefore independent of processes from the past. Note that the suppressed subscript of ε_X is $t-\tau$ for $\tau > 0$. Further, $E[Y^\top\mathbf{U}]$ becomes

$$E[Y^\top\mathbf{U}] = \left(\mathbf{B}^\top, c\mathbf{A}^\top\right) E[\mathbf{U}^\top\mathbf{U}] + cE[\varepsilon_X^\top\mathbf{U}] + \underbrace{E[\varepsilon_Y^\top\mathbf{U}]}_{=0}, \tag{A.130}$$

and

$$E[\mathbf{U}^\top X] = E[\mathbf{U}^\top\mathbf{U}]\begin{pmatrix}0\\\mathbf{A}\end{pmatrix} + E[\mathbf{U}^\top X]. \tag{A.131}$$

Then

$$\begin{aligned}&E[Y^\top\mathbf{U}](E[\mathbf{U}^\top\mathbf{U}])^{-1}E[\mathbf{U}^\top X] = \\&\left(\mathbf{B}^\top, c\mathbf{A}^\top\right) E[\mathbf{U}^\top\mathbf{U}]\begin{pmatrix}0\\\mathbf{A}\end{pmatrix} + cE[\varepsilon_X^\top\mathbf{U}]\begin{pmatrix}0\\\mathbf{A}\end{pmatrix} +\\&+ \left(\mathbf{B}^\top, c\mathbf{A}^\top\right) E[\mathbf{U}^\top\varepsilon_X] + cE[\varepsilon_X^\top\mathbf{U}](E[\mathbf{U}^\top\mathbf{U}])^{-1}E[\mathbf{U}^\top\varepsilon_X].\end{aligned} \tag{A.132}$$

Thus, many terms in $E[Y_\mathbf{U}^\top X_\mathbf{U}]$ cancel, and it remains

$$E[Y_\mathbf{U}^\top X_\mathbf{U}] = c\sigma_X^2 - c\underbrace{E[\varepsilon_X^\top\mathbf{U}](E[\mathbf{U}^\top\mathbf{U}])^{-1}E[\mathbf{U}^\top\varepsilon_X]}_{(\star)}. \tag{A.133}$$

Treating the inverse covariance in the $(\star)$-term with the matrix inversion lemma analogous to Eq. (A.107) and noting that

$$E[\varepsilon_X^\top\mathbf{U}] = \left(0, \varepsilon_X^\top\mathbf{W}\right), \tag{A.134}$$

because ε_X is independent from the parents $\mathbf{Z}$ of X, the $(\star)$-term becomes

$$(\star) = E[\varepsilon_X^\top\mathbf{W}]S_Z^{-1}E[\mathbf{W}^\top\varepsilon_X]. \tag{A.135}$$

S_Z^{-1} is again the inverted $N_Y \times N_Y$ matrix of the conditional variance of $\mathbf{W}$ given $\mathbf{Z}$,

$$S_Z = E[\mathbf{W}^\top \mathbf{W}] - E[\mathbf{W}^\top \mathbf{Z}](E[\mathbf{Z}^\top \mathbf{Z}])^{-1} E[\mathbf{Z}^\top \mathbf{W}]. \tag{A.136}$$

Along the same derivation the variances are evaluated. All together, the covariances and variances are simplified to

$$\begin{aligned} E[Y_\mathbf{U}^\top X_\mathbf{U}] &= c\sigma_X^2 \\ &\quad - cE[\varepsilon_X^\top \mathbf{W}] S_Z^{-1} E[\mathbf{W}^\top \varepsilon_X] \qquad \text{(A.137)} \\ E[Y_\mathbf{U}^\top Y_\mathbf{U}] &= \sigma_Y^2 + c^2\sigma_X^2 \\ &\quad - c^2 E[\varepsilon_X^\top \mathbf{W}] S_Z^{-1} E[\mathbf{W}^\top \varepsilon_X] \qquad \text{(A.138)} \\ E[X_\mathbf{U}^\top X_\mathbf{U}] &= \sigma_X^2 \\ &\quad - E[\varepsilon_X^\top \mathbf{W}] S_Z^{-1} E[\mathbf{W}^\top \varepsilon_X]. \qquad \text{(A.139)} \end{aligned}$$

The "sidepath" contribution $E[\varepsilon_X^\top \mathbf{W}]$ can be further analyzed as follows. Inserting t and τ again, the entries of the vector $E[\varepsilon_X^\top \mathbf{W}]$ can be written as

$$(E[\varepsilon_X^\top \mathbf{W}])_i = E[\varepsilon_{X,t-\tau} W^i_{t-g_i}], \tag{A.140}$$

A simple case where $E[\varepsilon_X^\top \mathbf{W}]$ is zero is given if $\forall i\ \tau < g_i$, i.e., all parents of Y are in the past of X. But it is interesting to further analyze more complex cases for $\tau \geq g_i$ for any i. Consider

$$\begin{aligned} E[\varepsilon_{X,t-\tau} W^i_{t-g_i}] &= E[W^i_{t+\tau-g_i} \varepsilon_{X,t}] \\ &= \underbrace{E[W^i_{t+\tau-g_i} X_t]}_{\Gamma_{W_i X}(\tau - g_i)} - \sum_{j=1}^{N_X} \Phi_{XZ_j}(h_j) \underbrace{E[W^i_{t+\tau-g_i} Z^j_{t-h_j}]}_{\Gamma_{W_i Z_j}(\tau + h_j - g_i)}. \end{aligned} \tag{A.141}$$

Analyzing $\Gamma_{W_i X}(\tau - g_i)$,

$$\Gamma_{W_i X}(\tau - g_i) = \sum_{n=0}^{\infty} \sum_{r=1}^{N} \sum_{s=1}^{N} \Psi_{W_i r}(n + \tau - g_i) \Sigma_{rs} \Psi_{Xs}(n), \tag{A.142}$$

the linear combination of paths in $\Psi_{Xs}(n)$ can be separated as they either all go through the parents of X or are emanating from X, i.e., are of length $n = 0$:

$$\Psi_{Xs}(n) = \delta_{X,s}\delta_{n,0} + \sum_{j=1}^{N_X} \Phi_{XZ_j}(h_j) \Psi_{Z_j s}(n - h_j) \tag{A.143}$$

resulting in

$$\Gamma_{W_i X}(\tau - g_i) =$$

$$= \sum_{n=0}^{\infty}\sum_{r=1}^{N}\sum_{s=1}^{N}\Psi_{W_i r}(n+\tau-g_i)\Sigma_{rs}\delta_{X,s}\delta_{n,0}+$$
$$+\sum_{n=0}^{\infty}\sum_{r=1}^{N}\sum_{s=1}^{N}\Psi_{W_i r}(n+\tau-g_i)\Sigma_{rs}\sum_{j=1}^{N_X}\Phi_{XZ_j}(h_j)\Psi_{Z_j s}(n-h_j) \tag{A.144}$$
$$= \sum_{r=1}^{N}\Psi_{W_i r}(\tau-g_i)\Sigma_{rX}+$$
$$+\sum_{j=1}^{N_X}\Phi_{XZ_j}(h_j)\sum_{n=0}^{\infty}\sum_{r=1}^{N}\sum_{s=1}^{N}\underbrace{\Psi_{W_i r}(n+\tau-g_i)\Sigma_{rs}\Psi_{Z_j s}(n-h_j)}_{\Psi_{W_i r}(n+\tau-g_i+h_j)\Sigma_{rs}\Psi_{Z_j s}(n)} \tag{A.145}$$
$$= \sum_{r=1}^{N}\Psi_{W_i r}(\tau-g_i)\Sigma_{rX}+\sum_{j=1}^{N_X}\Phi_{XZ_j}(h_j)\Gamma_{W_i Z_j}(\tau+h_j-g_i) \tag{A.146}$$

and thus

$$(E[\varepsilon_X^\top \mathbf{W}])_i = \sum_{r=1}^{N}\Psi_{W_i r}(\tau-g_i)\Sigma_{rX}. \tag{A.147}$$

$(E[\varepsilon_X^\top \mathbf{W}])_i$ is the linear combination of all paths of length $\tau - g_i$ emanating from X_t or $\mathbf{X}_{r,t}$ with $\Sigma_{rX} \neq 0$ to $W^i_{t+\tau-g_i}$.

For $\tau < g_i$, $\Psi(n<0) \equiv 0$ and thus for all i $(E[\varepsilon_X^\top \mathbf{W}])_i = 0$, confirming the first part of the theorem. But for all i with $\tau \geq g_i$, $(E[\varepsilon_X^\top \mathbf{W}])_i$ can still be zero if there are no such paths. If that holds for all i, the vector $E[\varepsilon_X^\top \mathbf{W}]$ is zero and the simple expression for MIT is obtained.

□

Appendix B.

Further climatological analyses

B.1. Stationarity analysis of Walker example

To further assess the stationarity of the Walker circulation example in Sect. 6.4, published in Runge et al. (2014), we conducted a sliding window analysis with windows of length 30 years (360 month samples) in steps of 3 years leading to 12 windows (albeit only 2 are non-overlapping). We used the same algorithm parameters and significance level as before. The results are shown in Fig. B.1. As discussed in the main section, the link WPAC $\rightarrow$ EPAC is significant only after 1970, but unchanged in the bivariate (Fig. B.1(a)) and trivariate (Fig. B.1(b)) analyses. Apart from the contemporaneous link between CPAC and WPAC, all other links are rather stationary. Note that the use of daily data would yield more precise lags, but also bring about the problem that a much larger number of significance tests has to be conducted.

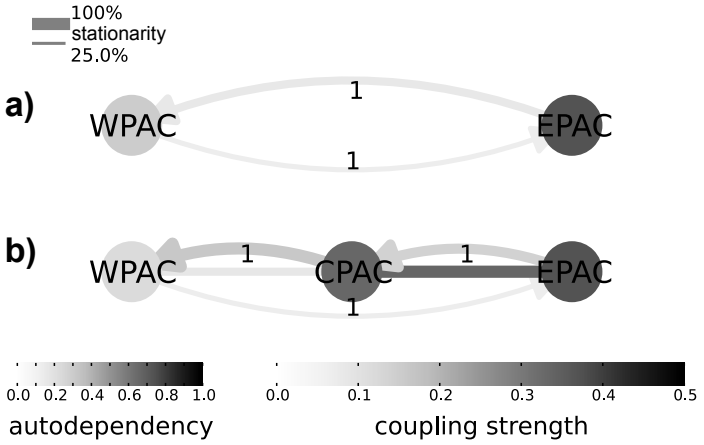


Figure B.1.: Ensemble statistics of sliding window analysis of the bi- and trivariate Walker circulation example. In this process graph the node color denotes the ensemble average auto-MIT lag-1 strength and the link color encodes the ensemble average MIT strength at the lag denoted in the link label. The link width, on the other hand, is proportional to the fraction of sliding windows with this link being significant.

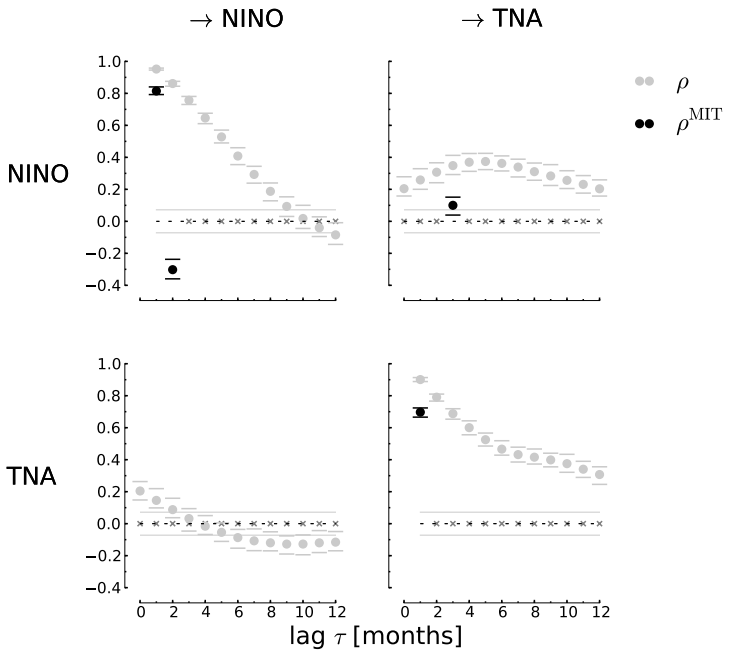


Figure B.2.: Correlations and partial correlations of Nino3.4 – TNA example. The matrix of lag functions shows the (auto-)correlations (light gray) and the value of MIT (black), where non-significant links are marked by gray crosses. The horizontal gray line denotes the two-sided 95%-significance level for the (auto-)correlations. The error bars mark the 90% confidence interval estimated from a bootstrap test. For example, the upper right plot shows the lagged cross correlation function $\rho(\mathrm{NINO}_{t-\tau}; \mathrm{TNA}_t)$ for $\tau \geq 0$ in light gray and the MIT value at the significant link "$\mathrm{NINO}_{t-3} \to \mathrm{TNA}_t$" in black. Note that for autocorrelations (on the diagonal) the zero-lag is not drawn.

B.2. Further example of Pacific – Atlantic interaction

As a further climatological application that shows the robustness of our results on the Pacific – Atlantic teleconnection studied in Sect. 6.3, we study two indices of monthly sea surface temperature anomalies (Rayner et al., 2003) for the period 1950 – 2012. These results are from Runge (2013). NINO is, again, the time series of the spatial average over the Nino3.4 region and TNA is the tropical North Atlantic index (Enfield et al., 1999) averaged over (15°–57.5°W, 5.5°N–23.5°N).

Figure B.2 shows the (partial) correlations. The time series graph was estimated using the PC algorithm with (two-sided) significance level $\alpha = 95\%$. The estimated time series graph is comprised of a coupling link "$\mathrm{NINO}_{t-3} \to \mathrm{TNA}_t$" and autodependency links at lag 1 and 2 in NINO and only at lag 1 in TNA. On the other hand, the auto- and cross correlation lag functions shown in grey feature significant links for a large range of lags with a maximum of the cross correlation lag function $\rho(\mathrm{NINO}_{t-\tau}; \mathrm{TNA}_t)$

at lag $\tau = 5$, indicating a shift also here as found also in Sect. 2.2.2 and 6.3. Also the cross correlation value $\rho = 0.35 \pm 0.05$ at lag $\tau = 3$ is significantly larger than $\rho^{\mathrm{MIT}}(\tau = 3) = 0.10 \pm 0.05$ (the "±" values correspond to the 90% confidence interval estimated from the bootstrap test described in Sect. 4.3.4).

The strong autodependency links with MIT values of (0.8, −0.3) for lags 1 and 2 in NINO and 0.7 for lag 1 in TNA explain these 'significant' cross correlation values at most lags, which, according to Eq. (A.10), are due to the common driver effect of past nodes (Fig. 2.3(b)) or the indirect causal effect due to intermediate lags (Fig. 2.3(a)). On the other hand, since there are no sidepaths here (at least among the two processes studied), the small MIT value reflects only the contributions from the coupling link and the residual's variances according to Eq. (A.121). The small value of MIT shows, that the actual coupling mechanism by which NINO influences TNA is quite weak, but due to strong autocorrelations the overall contribution to TNA's variance is larger becoming maximal in the peak at lag 5 consistent with Sect. 6.3.

B.3. Vertical interactions in the tropics

In the nonlinear framework, we analyze monthly air temperature anomalies in the tropics at two different altitudes in a NCEP/NCAR reanalysis dataset (Compo et al., 2006) showing results from Runge et al. (2012b). To investigate the upwelling of heat from the sea surface towards the upper troposphere in a height of about 12 km, we measure the coupling strength between the surface pressure level (X in Fig. B.3) and the 200 hPa pressure level (Y) for all tropical (latitudes between 30°S and 30°N) grid points.

First, we estimated the time series graph with conditional mutual information separately for each surface – troposphere pair at each grid point using a significance threshold estimated with the shuffle test at the $\alpha - 0.98$ level with $k = 100$. On average, we found the parents $\mathcal{P}_{X_t} = \{X_{t-1}\}$ and $\mathcal{P}_{Y_t} = \{Y_{t-1}\}$, i.e., lag-1 autodependencies, and the contemporaneous link "$X_t - Y_t$". With these parents, the spatial average of all lag functions of MIT in the left panel of Fig. B.3(a) shows the contemporaneous link "$X_t - Y_t$" as a significant peak, indicating that the time scale of the coupling is below the lag of one month. The MI, on the other hand, is significant for a wide range of lags, making an assessment of a physical coupling delay difficult. While the contemporaneous link cannot be interpreted as a directed coupling, we can still assess its strength. The MIT of a linear Gaussian process with the same time series graph is $I^{\mathrm{MIT}}_{X-Y} = \frac{1}{2} \log \left(\frac{\sigma_X^2 \sigma_Y^2}{\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2} \right)$, while MI additionally depends on the autodependency coefficients.

Figure B.3(b) shows a large (compared to the extra tropics) I^{MI}_{X-Y} all across the tropics. Significant I^{MIT}_{X-Y} values, on the other hand, are more confined and largest between 90°E and 170°W. Larger MIT values indicate a stronger coupling between the surface and upper tropospheric level in an area that actually corresponds to a region of strong upwelling in the Walker circulation (Lau and Yang, 2003). The difference between MI and MIT is largest in the Eastern Pacific where also the

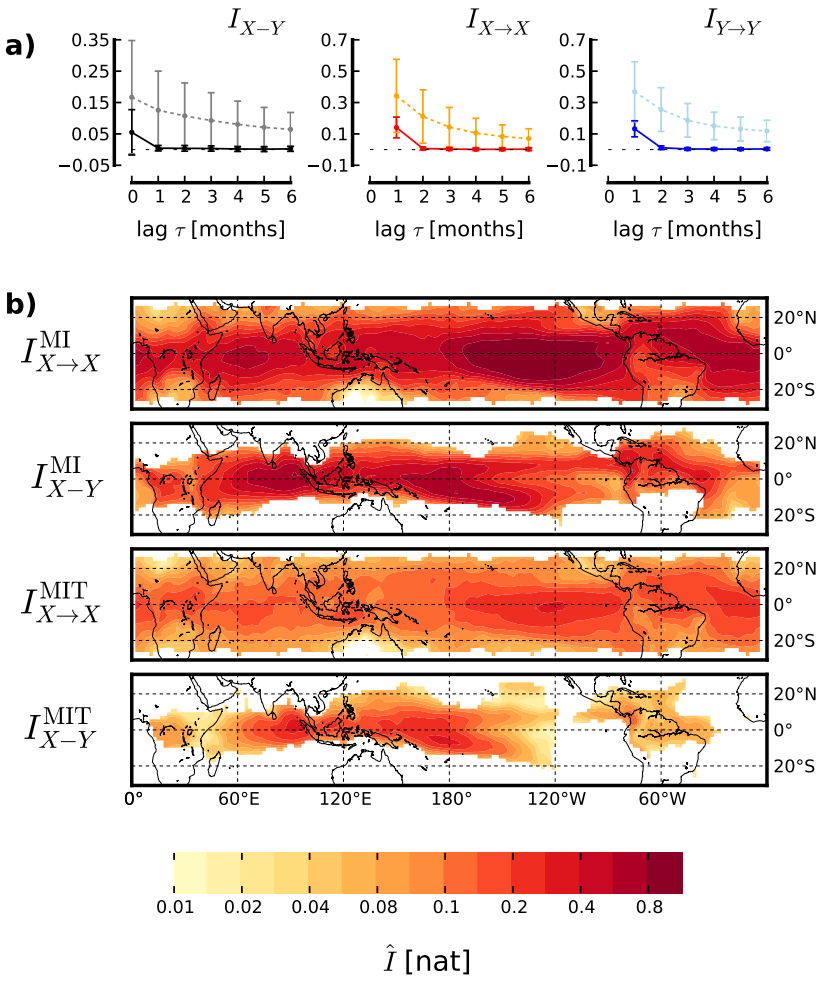


Figure B.3.: Analysis of air temperature anomalies at the surface (X) and the upper troposphere (Y), $T = 1008$ months (1927–2011). (a) Spatial average and standard deviation of coupling (left plot) and autodependency (middle plot for X, right plot for Y) lag functions for MI (dashed lines in light colors) and MIT (solid lines in dark colors). (b) Spatially resolved coupling strengths of the contemporaneous link "$X_t - Y_t$" and the autodependency "$X_{t-1} \to X_t$" for MI (upper two panels) and MIT (lower two panels). $I^{\mathrm{MI}}_{Y\to Y}$ and $I^{\mathrm{MIT}}_{Y\to Y}$ (not shown) are almost the same all across the tropics. For the contemporaneous link, values below the 98% significance level are in white. CMIs estimated with $k = 10$ here ($k = 100$ in the PC algorithm to determine parents).

increased autodependency in surface air temperatures is apparent ($I^{\mathrm{MIT}}_{X\to X}$). This strong persistence thus leads to a spurious increase in MI, which cannot differentiate the effects of increased autodependencies and increased contemporaneous coupling like MIT. With our measure of coupling strength we are, thus, able to infer a more

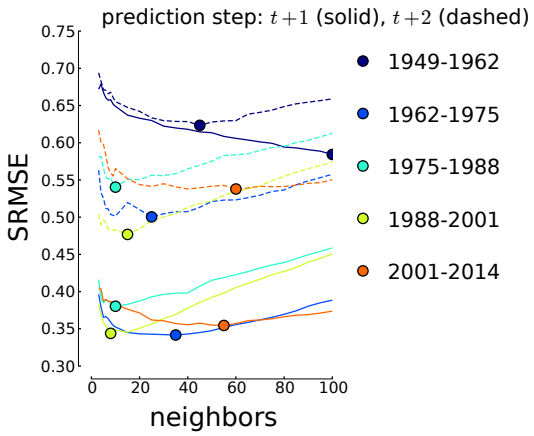


Figure B.4.: Robustness of prediction of the ENSO index Nino3.4 using Eq. (7.10) to nearest-neighbor parameter n. Shown is the standardized root mean squared error given by Eq. (7.12) evaluated using 5-fold cross-validation for the periods indicated in the legend for one (solid) and two months (dashed) ahead predictions plotted against the number of nearest-neighbors n used. The number with minimum out-of-sample error is marked by a colored dot. The number of predictors is shown in Fig. 7.2, estimated from the learning set with our heuristic criterion (7.8).

reasonable picture of the physical interactions in the Walker circulation complementing the analysis in Sect. 6.4. This further example underlines the importance of having a meaningfully interpretable coupling measure.

B.4. Robustness of prediction

For the nearest-neighbor prediction scheme discussed in Sect. 7.3, next to the number of predictors, also the number of nearest neighbors n used in Eq. (7.10) can affect the optimality of a prediction. In Fig. B.4, we show that our choice of $n = 10$ is rather robust for the example of predicting Nino3.4 in Sect. 7.3. For larger values of n, the prediction takes into account not only local information, but increasingly averages across the whole sample leading to worsened predictions.

Bibliography

Abarbanel, H. D. I., R. Brown, and J. B. Kadtke (1990). "Prediction in chaotic nonlinear systems: Methods for time series with broadband Fourier spectra". In: *Physical Review A* 41.4 (cit. on p. 157).

Abramson, N. (1963). *Information theory and coding*. New York, NY: McGraw-Hill (cit. on p. 42).

Adler, R. F., G. J. Huffman, A. Chang, R. Ferraro, P.-P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin, and E. Nelkin (2003). "The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-Present)". EN. In: *Journal of Hydrometeorology* 4.6, pp. 1147–1167 (cit. on p. 128).

Alexander, M. A., N. C. Lau, and J. D. Scott (2004). "Broadening the Atmospheric Bridge Paradigm: ENSO Teleconnections to the Tropical West Pacific-Indian Oceans Over the Seasonal Cycle and to the North Pacific in Summer". In: *Earth's Climate* 147, pp. 85–104 (cit. on p. 148).

Allahverdyan, A. E., D. Janzing, and G. Mahler (2009). "Thermodynamic efficiency of information and heat flow". In: *Journal of Statistical Mechanics: Theory and Experiment*, P09011 (cit. on pp. 37, 114).

Alparslan, A. K., M. Sayar, and A. R. Atilgan (1998). "State-space prediction model for chaotic time series". In: *Physical Review E* 58.2, pp. 2640–2643 (cit. on p. 157).

Amarasingham, A., M. T. Harrison, N. G. Hatsopoulos, and S. Geman (2012). "Conditional modeling and the jitter method of spike resampling". In: *Journal of Neurophysiology* 107, pp. 517–531 (cit. on pp. 5, 171).

Amblard, P. O. and O. J. J. Michel (2011). "On directed information theory and Granger causality graphs". In: *Journal of Computational Neuroscience* 30.1, pp. 7–16 (cit. on p. 19).

– (2012). "The relation between Granger causality and directed information theory: a review". In: *Entropy* 15.1, pp. 113–143 (cit. on p. 16).

Ancona, N., D. Marinazzo, and S. Stramaglia (2004). "Radial basis function approach to nonlinear Granger causality of time series". In: *Physical Review E* 70, p. 056221 (cit. on p. 17).

Ansell, T. J., P. D. Jones, R. J. Allan, D. Lister, D. E. Parker, M. Brunet, A. Moberg, J. Jacobeit, P. Brohan, N. A. Rayner, and Others (2010). "Daily mean sea level pressure reconstructions for the European-North Atlantic region for the period 1850-2003". In: *Journal of Climate* 19, pp. 2717–2742 (cit. on p. 126).

Arenas, A., A. Diaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou (2008). "Synchronization in complex networks". In: *Physics Reports* 469.3, pp. 93–153 (cit. on pp. 12, 18).

Ashok, K., Z. Guan, N. H. Saji, and T. Yamagata (2004). "Individual and combined influences of ENSO and the Indian Ocean dipole on the Indian summer monsoon." In: *Journal of Climate* 17.16, pp. 3141–3155 (cit. on p. 153).

Ay, N. and D. Polani (2008). "Information flows in causal networks". In: *Advances in complex systems* 11.1, pp. 17–42 (cit. on pp. 35, 113).

Bahraminasab, A., F. Ghasemi, A. Stefanovska, P. V. E. McClintock, and H. Kantz (2008). "Direction of coupling from phases of interacting oscillators: A permutation information approach". In: *Physical Review Letters* 100.8, p. 84101 (cit. on p. 19).

Balasis, G., R. V. Donner, S. M. Potirakis, J. Runge, C. Papadimitriou, I. Daglis, K. Eftaxis, and J. Kurths (2013). "Statistical Mechanics and Information-Theoretic Perspectives on Complexity in the Earth System". In: *Entropy* 15.11, pp. 4844–4888 (cit. on pp. 134, 170).

Barnett, L., A. B. Barrett, and A. K. Seth (2009). "Granger causality and transfer entropy are equivalent for Gaussian variables". In: *Physical Review Letters* 103, p. 238701 (cit. on pp. 20, 46).

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt (2012). "Skill of Real-Time Seasonal ENSO Model Predictions during 2002-11: Is Our Capability Increasing?" In: *Bulletin of the American Meteorological Society* 93.5, pp. 631–651 (cit. on pp. 157, 165).

Barrett, A. B., L. Barnett, and A. K. Seth (2010). "Multivariate Granger causality and generalized variance". In: *Physical Review E* 81.4, p. 41907 (cit. on p. 17).

Beirlant, J. and E. J. Dudewicz (1997). "Nonparametric entropy estimation: An overview". In: *International Journal of Mathematical and Statistical Science* 6.1, pp. 1–14 (cit. on p. 60).

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press (cit. on p. 4).

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B* 57.1, pp. 289–300 (cit. on p. 140).

Bergsma, W. (2004). "Testing conditional independence for continuous random variables". In: *Eurandom technical report* 48, pp. 1–19 (cit. on p. 66).

Bialonski, S., M. T. Horstmann, and K. Lehnertz (2010). "From brain to earth and climate systems: Small-world interaction networks or not?" In: *Chaos* 20.1, p. 13134 (cit. on p. 118).

Bjerknes, J. (1969). "Atmospheric teleconnections from the equatorial pacific". In: *Monthly Weather Review* 97.3, pp. 163–172 (cit. on p. 136).

Blinowska, K. and M. Kaminski (2013). "Functional Brain Networks: Random, Small World or Deterministic?" In: *PloS one* 8.10, e78763 (cit. on p. 118).

Boccaletti, S., J. Kurths, and G. Osipov (2002). "The synchronization of chaotic systems". In: *Physics Reports* 366.1-2, pp. 1–101 (cit. on pp. 12, 18).

Boccaletti, S, V Latora, Y Moreno, M Chavez, and D. U. Hwang (2006). "Complex networks: Structure and dynamics". In: *Physics Reports* 424.4-5, pp. 175–308 (cit. on p. 118).

Boltzmann, L. (1872). "Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen". In: *Sitzungsberichte der Akademie der Wissenschaften Wien* 2.66 (cit. on p. 37).

Böttcher, A., B. Silbermann, and A. Karlovich (2006). *Analysis of Toeplitz operators*. Berlin Heidelberg: Springer (cit. on pp. 94, 183).

Bouezmarni, T., J. Rombouts, and A Taamouti (2009). "A nonparametric copula based test for conditional independence with applications to Granger causality". In: *UC3M Working Paper. Economic Series* 9.34 (cit. on pp. 21, 64, 66).

Breitenbach, S. F. M., K. Rehfeld, B. Goswami, J. U. L. Baldini, H. E. Ridley, D. J. Kennett, K. M. Prufer, V. V. Aquino, Y. Asmerom, V. J. Polyak, H. Cheng, J. Kurths, and N. Marwan (2012). "COnstructing Proxy Records from Age models (COPRA)". In: *Climate of the Past* 8.5, pp. 1765–1779 (cit. on p. 30).

Brockwell, P. J. and R. A. Davis (2002). *Introduction to time series and forecasting*. 2nd. New York: Springer (cit. on pp. 116, 161).

Brockwell, P. and R. Davis (2009). *Time series: theory and methods*. New York: Springer (cit. on pp. 17, 101, 105, 178, 179).

Bullmore, E. and O. Sporns (2009). "Complex brain networks: graph theoretical analysis of structural and functional systems". In: *Nature Reviews Neuroscience* 10, pp. 186–198 (cit. on pp. 2, 3, 11, 118).

Cane, M. A. (2005). "The evolution of El Niño, past and future". In: *Earth and Planetary Science Letters* 230.3-4, pp. 227–240 (cit. on pp. 2, 125).

Cane, M. A., S. E. Zebiak, and S. C. Dolan (1986). "Experimental forecasts of El Niño". In: *Nature* 321, pp. 827–832 (cit. on p. 157).

Casdagli, M., S. Eubank, J. Farmer, and J. Gibson (1991). "State space reconstruction in the presence of noise". In: *Physica D: Nonlinear Phenomena* 51.1-3, pp. 52–98 (cit. on p. 18).

Chang, P., Y. Fang, R. Saravanan, L. Ji, and H. Seidel (2006). "The cause of the fragile relationship between the Pacific El Niño and the Atlantic Niño." In: *Nature* 443.7109, pp. 324–8 (cit. on p. 133).

Chatfield, C. (2013). *The analysis of time series: an introduction*. London: Chapman & Hall/CRC (cit. on pp. 14, 74).

Chen, Y., G. Rangarajan, J. Feng, and M. Ding (2004). "Analyzing multiple nonlinear time series with extended Granger causality". In: *Physics Letters A* 324.1, pp. 26–35 (cit. on pp. 18, 35).

Chen, Y., S. L. Bressler, and M. Ding (2006). "Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data". In: *Journal of neuroscience methods* 150.2, pp. 228–237 (cit. on p. 17).

Chu, T. and C. Glymour (2008). "Search for additive nonlinear time series causal models". In: *The Journal of Machine Learning Research* 9.5, pp. 967–991 (cit. on p. 24).

Clausius, R. (1862). "Über die Wärmeleitung gasförmiger Körper". In: *Annalen der Physik* 125.7 (cit. on p. 37).

Compo, G., J. S. Whitaker, and P. Sardeshmukh (2006). "Feasibility of a 100-year reanalysis using only surface pressure data". In: *Bulletin of the American Meteorological Society* 87.2, pp. 175–190 (cit. on p. 206).

Cover, T. M. and J. A. Thomas (2006). *Elements of information theory*. Hoboken: John Wiley & Sons (cit. on pp. 19, 40–42, 113, 178).

Crutchfield, J. P. and C. R. Shalizi (1999). "Thermodynamic depth of causal states: Objective complexity via minimal representations". In: *Physical Review E* 59.275 (cit. on pp. 37, 114).

Dahlhaus, R. (2000). "Graphical interaction models for multivariate time series". In: *Metrika* 51.2, pp. 157–172 (cit. on p. 22).

Dahlhaus, R. and M. Eichler (2003). "Causality and graphical models in time series analysis". In: *Oxford Statistical Science Series*, pp. 115–137 (cit. on p. 22).

Daniusis, P., D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf (2012). "Inferring deterministic causal relations". In: *arXiv preprint* 1203.3475 [cs.LG] (cit. on p. 31).

Darbellay, G. and I. Vajda (1999). "Estimation of the information by an adaptive partitioning of the observation space". In: *IEEE Transactions on Information Theory* 45.4, pp. 1315–1321 (cit. on p. 60).

Deshpande, G., P. Santhanam, and X. Hu (2011). "Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data." In: *NeuroImage* 54.2, pp. 1043–52 (cit. on p. 118).

Detto, M., A. Molini, G. Katul, P. Stoy, S. Palmroth, and D. Baldocchi (2012). "Causality and persistence in ecological systems: a nonparametric spectral granger causality approach." In: *The American naturalist* 179.4, pp. 524–35 (cit. on p. 17).

Ding, M., Y. Chen, and S. L. Bressler (2006). "Granger Causality: Basic Theory and Application to Neuroscience". In: *Handbook of Time Series Analysis*. Ed. by B. Schelter, M. Winterhalder, and J. Timmer. Wiley-VCH Verlage, pp. 451–474 (cit. on p. 16).

Dobrushin, R. L. (1958). "A simplified method of experimentally evaluating the entropy of a stationary sequence". In: *Theory of Probability & Its Applications* 3.4 (cit. on p. 60).

Donges, J. F. (2012). "Functional network macroscopes for probing past and present Earth system dynamics". PhD thesis. Humboldt University (cit. on pp. 12, 118).

Donges, J. F., Y. Zou, N. Marwan, and J. Kurths (2009a). "Complex networks in climate dynamics". In: *European Physical Journal Special Topics* 174.1, pp. 157–179 (cit. on pp. 2, 3, 11, 12, 118, 143, 154, 173).

Donges, J. F., Y. Zou, N. Marwan, and J. Kurths (2009b). "The backbone of the climate network". In: *Europhysics Letters* 87.4 (cit. on p. 12).

Donges, J. F., H. C. H. Schultz, N. Marwan, Y. Zou, and J. Kurths (2011). "Investigating the topology of interacting networks – Theory and application to coupled climate subnetworks". In: *European Physical Journal B* 84.4, pp. 635–652 (cit. on p. 12).

Dufour, J. M. and E. Renault (1998). "Short run and long run causality in time series: theory". In: *Econometrica* 66.5, pp. 1099–1125 (cit. on p. 159).

Ebert-Uphoff, I. and Y. Deng (2012a). "Causal discovery for climate research using graphical models". In: *Journal of Climate* 25.17, pp. 5648–5665 (cit. on pp. 17, 126, 154).

Ebert-Uphoff, I. and Y. Deng (2012b). "A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer". In: *Geophysical Research Letters* 39.19, p. L19701 (cit. on pp. 4, 17, 126, 154, 172).

Ebisuzaki, W. (1997). "A method to estimate the statistical significance of a correlation when the data are serially correlated". In: *Journal of Climate* 10.9, pp. 2147–2153 (cit. on p. 72).

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap*. New York: Springer (cit. on p. 77).

Eichler, M. (2006). "On the evaluation of information flow in multivariate systems by the directed transfer function". In: *Biological cybernetics* 94.6, pp. 469–482 (cit. on p. 17).

– (2012). "Graphical modelling of multivariate time series". In: *Probability Theory and Related Fields* 153.1, pp. 233–268 (cit. on pp. 22–27).

Eichler, M. (2005). "A graphical approach for evaluating effective connectivity in neural systems." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.1457, pp. 953–67 (cit. on pp. 22, 31).

Einstein, A. (1905). "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt". In: *Annalen der Physik* 322.6, pp. 132–148 (cit. on p. 114).

Eisenstein, E., I. Kanter, D. A. Kessler, and W. Kinzel (1995). "Generation and prediction of time series by a neural network". In: *Physical Review Letters* 74.1, pp. 1–4 (cit. on p. 157).

Enfie, D. B. and A. Mayer (1997). "Tropical Atlantic sea surface temperature variability relation". In: *Journal of Geophysical Research: Oceans (1978–2012)* 102.C1, pp. 929–945 (cit. on p. 133).

Enfield, D. B., A. M. Mestas-Nuñez, D. A. Mayer, and L. Cid-Serrano (1999). "How ubiquitous is the dipole relationship in tropical Atlantic sea surface temperatures?" In: *Journal of Geophysical Research* 104.C4, p. 7841 (cit. on p. 205).

Faes, L., A. Porta, and G. Nollo (2008). "Mutual nonlinear prediction as a tool to evaluate coupling strength and directionality in bivariate time series: Comparison among different strategies based on k nearest neighbors". In: *Physical Review E* 78.2, p. 026201 (cit. on p. 18).

Faes, L., G. Nollo, and A. Porta (2011). "Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique". In: *Physical Review E* 83.5 (cit. on p. 18).

– (2013). "Compensated Transfer Entropy as a Tool for Reliably Estimating Information Transfer in Physiological Time Series". In: *Entropy* 15.1, pp. 198–219 (cit. on p. 19).

Fan, J. (2003). *Nonlinear time series: nonparametric and parametric methods*. New York: Springer (cit. on p. 16).

Farmer, J. D. and J. J. Sidorowich (1987). "Predicting chaotic time series". In: *Physical review letters* 59, pp. 845–848 (cit. on pp. 157, 161).

Feldhoff, J. H., R. V. Donner, J. F. Donges, N. Marwan, and J. Kurths (2012). "Geometric detection of coupling directions by means of inter-system recurrence networks". In: *Physics Letters A* 376.46, pp. 3504–3513 (cit. on p. 18).

Fisher, R. A. (1924). "The Distribution of the Partial Correlation Coefficient." In: *Metron* 3, pp. 329–332 (cit. on p. 74).

Florens, J. P. and M. Mouchart (1982). "A note on noncausality". In: *Econometrica: Journal of the Econometric Society* 50.3, pp. 583–591 (cit. on p. 22).

Frankignoul, C. and K. Hasselmann (1977). "Stochastic climate models, part II application to sea-surface temperature anomalies and thermocline variability". In: *Tellus* 29.1977, pp. 289–305 (cit. on p. 116).

Frenzel, S. and B. Pompe (2007). "Partial Mutual Information for Coupling Analysis of Multivariate Time Series". In: *Physical Review Letters* 99.20, p. 204101 (cit. on pp. 5, 20, 59–63, 68).

Friedman, J., T. Hastie, and R. Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (cit. on pp. 68, 138).

Gebelein, H. (1941). "Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung". In: *Zeitschrift für Angewandte Mathematik und Mechanik* 21.6, pp. 364–379 (cit. on p. 35).

Geweke, J. (1984). "Inference and causality in economic time series models". In: *Handbook of Econometrics*. Ed. by Z. Grilichest and M. D. Intriligator. 2nd ed. Amsterdam: Elsevier (cit. on p. 16).

Ghosh, M., N. Reid, and D. A. Fraser (2010). "Ancillary statistics: a review". In: *Statistica Sinica* 20.4, p. 1309 (cit. on p. 106).

Giannini, A., Y. Kushnir, and M. A. Cane (2000). "Interannual Variability of Caribbean Rainfall, ENSO, and the Atlantic Ocean". In: *Journal of Climate* 13.2, pp. 297–311 (cit. on p. 133).

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics Developed with Especial Reference to the Rational Fondation of Thermodynamics*. Cambridge: Cambridge University Press (cit. on p. 37).

Gibson, J. F., J. D. Farmer, M. Casdagli, and S. Eubank (1992). "An analytic approach to practical state space reconstruction". In: *Physica D* 57.1-2, pp. 1–30 (cit. on p. 18).

Giona, M., F. Lentini, and V. Cimagalli (1991). "Functional reconstruction and local prediction of chaotic time series". In: *Physical Review A* 44.6, pp. 3496–3502 (cit. on p. 157).

Golan, A. (2002). "Information and Entropy Econometrics – Editor's View". In: *Journal of Econometrics* 107.1-2, pp. 1–15 (cit. on p. 37).

Goria, M. N. and N. N. Leonenko (2005). "A new class of random vector entropy estimators and its applications in testing statistical hypotheses". In: *Nonparametric Statistics* 17.3, pp. 277–297 (cit. on p. 61).

Gourévitch, B., R. Le Bouquin-Jeannès, and G. Faucon (2006). "Linear and nonlinear causality between signals: methods, examples and neurophysiological applications". In: *Biological Cybernetics* 95.4, pp. 349–369 (cit. on p. 16).

Gozolchiani, A., K. Yamasaki, O. Gazit, and S. Havlin (2008). "Pattern of climate network blinking links follows El Niño events". In: *Europhysics Letters* 83.2, p. 28005 (cit. on p. 13).

Granger, C. W. J. (1969). "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: Journal of the Econometric Society* 37.3, pp. 424–438 (cit. on pp. 4, 16, 19, 24).

– (1988). "Some recent development in a concept of causality". In: *Journal of econometrics* 39.1-2, pp. 199–211 (cit. on pp. 4, 16).

Granger, C. W. J. and J.-L. Lin (1994). "Using the mutual information coefficient to identify lags in nonlinear models". In: *Journal of Time Series Analysis* 15.4, pp. 371–384 (cit. on pp. 14, 88, 93).

Groth, A. (2001). "Das Prinzip der maximalen Transinformation zur statistischen Modellierung". Diploma thesis. Ernst-Moritz-Arndt-Universität Greifswald (cit. on pp. 157, 158, 160, 161).

Groth, A. and M. Ghil (2011). "Multivariate singular spectrum analysis and the road to phase synchronization". In: *Physical Review E* 84, p. 036206 (cit. on p. 12).

Gu, G. and R. F. Adler (2011). "Precipitation and Temperature Variations on the Interannual Time Scale: Assessing the Impact of ENSO and Volcanic Eruptions". EN. In: *Journal of Climate* 24.9, pp. 2258–2270 (cit. on pp. 13, 115).

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press (cit. on p. 16).

Hashizume, M., T. Terao, and N. Minakawa (2009). "The Indian Ocean Dipole and malaria risk in the highlands of western Kenya." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.6, pp. 1857–62 (cit. on p. 13).

Hastie, T. and R. Tibshirani (1986). *Generalized additive models*. Monographs. Chapman & Hall/CRC (cit. on pp. 17, 81, 107).

Hiemstra, C. and J. Jones (1994). "Testing for linear and nonlinear Granger causality in the stock price-volume relation". In: *Journal of Finance* 49.5, pp. 1639–1664 (cit. on p. 16).

Hlaváčková-Schindler, K., M. Paluš, M. Vejmelka, and J. Bhattacharya (2007). "Causality detection based on information-theoretic approaches in time series analysis". In: *Physics Reports* 441.1, pp. 1–46 (cit. on pp. 19, 60, 61).

Hlinka, J., D. Hartman, and M. Paluš (2012). "Small-world topology of functional connectivity in randomly connected dynamical systems". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22.3 (cit. on p. 118).

Hlinka, J., D. Hartman, M. Vejmelka, J. Runge, N. Marwan, J. Kurths, and M. Paluš (2013). "Reliability of Inference of Directed Climate Networks Using Conditional Mutual Information". In: *Entropy* 15.6, pp. 2023–2045 (cit. on pp. 19, 118, 139, 170).

Hosking, J. S., M. R. Russo, P. Braesicke, and J. A. Pyle (2012). "Tropical convective transport and the Walker circulation". In: *Atmospheric Chemistry and Physics* 12.20, pp. 12229–12244 (cit. on p. 136).

Hsiao, C. (1982). "Autoregressive modeling and causal ordering of economic variables". In: *Journal of Economic Dynamics and Control* 4, pp. 243–259 (cit. on p. 159).

Huang, F., S. Zhou, S. Zhang, H. Wang, and L. Tang (2011). "Temporal correlation analysis between malaria and meteorological factors in Motuo County, Tibet." In: *Malaria Journal* 10, p. 54 (cit. on p. 116).

Hyvärinen, A., S. Shimizu, and P. O. Hoyer (2008). "Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity". In: *Proceedings of the 25th international conference on Machine learning*, pp. 424–431 (cit. on p. 17).

Jachan, M., K. Henschel, J. Nawrath, A. Schad, J. Timmer, and B. Schelter (2009). "Inferring direct directed-information flow from multivariate nonlinear time series". In: *Physical Review E* 80.1, pp. 1–5 (cit. on p. 35).

Jakulin, A. and I. Bratko (2003). *Analyzing attribute dependencies*. New York: Springer (cit. on p. 43).

Janzing, D., J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf (2012). "Information-geometric approach to inferring causal directions". In: *Artificial Intelligence* 108, pp. 1–31 (cit. on p. 31).

Janzing, D., D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf (2013). "Quantifying causal influences". In: *The Annals of Statistics* 41.5, pp. 2324–2358 (cit. on pp. 35, 75, 113, 170).

Jaynes, E. T. (1957). "Information theory and statistical mechanics". In: *Physical Review* 108.4, p. 620 (cit. on pp. 37, 114).

Jin, F. F. (1997). "An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model". In: *Journal of the Atmospheric Sciences* 54.7, pp. 811–829 (cit. on pp. 2, 125).

Kaiser, A. and T. Schreiber (2002). "Information transfer in continuous processes". In: *Physica D: Nonlinear Phenomena* 166.1-2, pp. 43–62 (cit. on p. 19).

Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* 23.3, pp. 187–200 (cit. on pp. 138, 140).

Kalisch, M. (2007). "Estimating high-dimensional directed acyclic graphs with the PC-algorithm". In: *The Journal of Machine Learning Research* 8, pp. 613–636 (cit. on p. 30).

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph (1996). "The NCEP/NCAR 40-Year Reanalysis Project". EN. In: *Bulletin of the American Meteorological Society* 77.3, pp. 437–471 (cit. on pp. 11, 128, 138).

Kantz, H. and T. Schreiber (2003). *Nonlinear Time Series Analysis*. Cambridge: Cambridge University Press, pp. 27–43 (cit. on pp. 18, 169).

Kaufmann, R. K. and D. I. Stern (1997). "Evidence for human influence on climate from hemispheric temperature relations". In: *Nature* 388.6637, pp. 39–44 (cit. on p. 17).

Khan, S., S. Bandyopadhyay, A. Ganguly, S. Saigal, D. Erickson, V. Protopopescu, and G. Ostrouchov (2007). "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data". In: *Physical Review E* 76.2, p. 026209 (cit. on p. 62).

Klein, S. A., B. J. Soden, and N. C. Lau (1999). "Remote sea surface temperature variations during ENSO: Evidence for a tropical atmospheric bridge". In: *Journal of Climate* 12.4, pp. 917–932 (cit. on p. 13).

Kozachenko, L. F. and N. N. Leonenko (1987). "Sample estimate of the entropy of a random vector". In: *Problemy Peredachi Informatsii* 23.2, pp. 9–16 (cit. on pp. 60, 61).

Kramer, G. (1998). "Directed information for channels with feedback". PhD Thesis. ETH Zürich (cit. on p. 19).

Kraskov, A., H. Stögbauer, and P. Grassberger (2004). "Estimating mutual information". In: *Physical Review E* 69.6, p. 066138 (cit. on pp. 60, 62, 70, 72).

Kuehn, C (2011). "A mathematical framework for critical transitions: Bifurcations, fast-slow systems and stochastic dynamics". In: *Physica D* 240.12, pp. 1020–1035 (cit. on pp. 153, 154).

Kugiumtzis, D. (2013). "Direct-coupling information measure from nonuniform embedding". In: *Physical Review E* 87.6, p. 062918 (cit. on p. 19).

Kullback, S. and R. A. Leibler (1951). "On information and sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86 (cit. on p. 40).

Kumar, K. K., B. Rajagopalan, and M. A. Cane (1999). "On the Weakening Relationship Between the Indian Monsoon and ENSO". In: *Science* 284.5423, pp. 2156–2159 (cit. on pp. 151–153).

Kumar, K. K., B. Rajagopalan, M. Hoerling, G. Bates, and M. Cane (2006). "Unraveling the mystery of Indian monsoon failure during El Niño ." In: *Science (New York, N.Y.)* 314.5796, pp. 115–9 (cit. on p. 152).

Lanzante, J. R. (1996). "Lag relationships involving tropical sea surface temperatures". In: *Journal of climate* 9.10, pp. 2568–2578 (cit. on pp. 13, 15).

Latif, M., D. Anderson, T. Barnett, M. A. Cane, R. Kleeman, A. Leetmaa, J. O'Brien, A. Rosati, and E. Schneider (1998). "A review of the predictability and prediction of ENSO". In: *Journal of Geophysical Research: Oceans (1978–2012)* 103.C7 (cit. on p. 157).

Latora, V. and M. Marchiori (2001). "Efficient behavior of small-world networks". In: *Physical Review Letters* 87.19, p. 198701 (cit. on pp. 3, 33).

Lau, K. M. and S. Yang (2003). "Walker circulation". In: *Encyclopedia of Atmospheric Sciences*, pp. 2505–2510 (cit. on pp. 148, 206).

Lauritzen, S. L. (1996). *Graphical Models*. Vol. 16. Oxford: Clarendon Press (cit. on pp. 21, 22, 26).

Lenton, T. M., H. Held, J. W. Hall, E. Kriegler, W. Lucht, S. Rahmstorf, and H.-J. Schellnhuber (2008). "Tipping elements in the Earth's climate system". In:

Proceedings of the National Academy of Sciences of the United States of America 105.6, pp. 1786–1793 (cit. on pp. 3, 152).

Leonenko, N. N., L. Pronzato, and V. Savani (2008). "A class of Rényi information estimators for multidimensional densities". In: *The Annals of Statistics* 36.5, pp. 2153–2182 (cit. on p. 61).

Leydesdorff, L. and Y. Sun (2009). "National and international dimensions of the Triple Helix in Japan: University-industry-government versus international coauthorship relations". In: *Journal of the American Society for Information Science and Technology* 60.4, pp. 778–788 (cit. on p. 42).

Liang, X. S. (2008). "Information flow within stochastic dynamical systems". In: *Physical Review E* 78.3, p. 031113 (cit. on p. 20).

– (2013). "The Liang-Kleeman Information Flow: Theory and Applications". In: *Entropy* 15.1, pp. 327–360 (cit. on pp. 20, 114, 170).

Liang, X. S. and R. Kleeman (2007a). "A rigorous formalism of information transfer between dynamical system components. II. Continuous flow". In: *Physica D: Nonlinear Phenomena* 227.2, pp. 173–182 (cit. on p. 20).

Liang, X. S. and R. Kleeman (2007b). "A rigorous formalism of information transfer between dynamical system components. I. Discrete mapping". In: *Physica D: Nonlinear Phenomena* 231.1, pp. 1–9 (cit. on p. 20).

Liang, X. S., R. Kleeman, and X. San Liang (2005). "Information transfer between dynamical systems components". In: *Physical Review Letters* 95.24, pp. 244101–1 (cit. on p. 20).

Liao, W., J. Ding, D. Marinazzo, Q. Xu, and Z. Wang (2011). "Small-world directed networks in the human brain: multivariate Granger causality analysis of resting-state fMRI". In: *Neuroimage* 54.4, pp. 2683–2694 (cit. on p. 118).

Lindley, D. V. (1956). "On a measure of the information provided by an experiment". In: *The Annals of Mathematical Statistics* 27.4, pp. 986–1005 (cit. on p. 37).

Lopez-Paz, D., P. Hennig, and B. Schölkopf (2013). "The Randomized Dependence Coefficient". In: *preprint arXiv:* 1304.7717, pp. 1–9 (cit. on p. 35).

Ludescher, J., A. Gozolchiani, M. I. Bogachev, A. Bunde, S. Havlin, and H. J. Schellnhuber (2013). "Improved El Niño forecasting by cooperativity detection". In: *Proceedings of the National Academy of Sciences* 110.29, pp. 11742–11745 (cit. on pp. 157, 165).

– (2014). "Very early warning of next El Niño". In: *Proceedings of the National Academy of Sciences* 111.6, pp. 2064–2066 (cit. on pp. 157, 165).

Majda, A. J. and J. Harlim (2007). "Information flow between subspaces of complex dynamical systems". In: *Proceedings of the National Academy of Sciences* 104.23, pp. 9558–9563 (cit. on p. 20).

Malik, N., B. Bookhagen, N. Marwan, and J. Kurths (2012). "Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks". In: *Climate Dynamics* 39.3-4, pp. 971–987 (cit. on p. 13).

Maraun, D. and J. Kurths (2005). "Epochs of phase coherence between El Niño/Southern Oscillation and Indian monsoon". In: *Geophysical Research Letters* 32.15, p. L15709 (cit. on pp. 18, 152).

Marinazzo, D., M. Pellicoro, and S. Stramaglia (2008). "Kernel method for nonlinear Granger causality". In: *Physical Review Letters* 100.14, p. 144103 (cit. on pp. 17, 35).

Marko, H. (1973). "The bidirectional communication theory-a generalization of information theory". In: *Communications, IEEE Transactions on [legacy, pre-1988]* 21.12, pp. 1345–1351 (cit. on p. 19).

Marwan, N., M. Carmen Romano, M. Thiel, and J. Kurths (2007). "Recurrence plots for the analysis of complex systems". In: *Physics Reports* 438.5-6, pp. 237–329 (cit. on p. 18).

Massey, J. (1990). "Causality, feedback and directed information". In: *Prococeedings of the International Symposium on Information Theory and its Applications*, pp. 303–305 (cit. on p. 19).

McGill, W. J. (1954). "Multivariate information transmission". In: *Psychometrika* 19.2, pp. 97–116 (cit. on p. 43).

Meinshausen, N. and P. Bühlmann (2006). "High-dimensional graphs and variable selection with the lasso". In: *The Annals of Statistics* 34.3, pp. 1436–1462 (cit. on p. 138).

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon (2002). "Network motifs: simple building blocks of complex networks." In: *Science* 298.5594, pp. 824–7 (cit. on p. 170).

Mokhov, I. I., D. A. Smirnov, P. I. Nakonechny, S. S. Kozlenko, E. P. Seleznev, and J. Kurths (2011). "Alternating mutual influence of El-Niño/Southern Oscillation and Indian monsoon". In: *Geophysical Research Letters* 38, pp. 2–6 (cit. on p. 152).

Mudelsee, M. (2010). *Climate time series analysis: Classical statistical and bootstrap methods*. Vol. 42. Atmospheric and Oceanographic Sciences Library. Dordrecht: Springer (cit. on pp. 72, 75).

Nawrath, J., M. Romano, M. Thiel, I. Kiss, M. Wickramasinghe, J. Timmer, J. Kurths, and B. Schelter (2010). "Distinguishing Direct from Indirect Interactions in Oscillatory Networks with Multiple Time Scales". In: *Physical Review Letters* 104.3, p. 38701 (cit. on p. 17).

Neureither, L. (2013). "Nonparametric Estimation of Entropy and its applications". Master Thesis. Freie Universität Berlin (cit. on p. 61).

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford: Oxford University Press (cit. on pp. 3, 12, 118).

Neyman, J. and E. L. Scott (1948). "Consistent estimates based on partially consistent observations". In: *Econometrica* 16.1 (cit. on pp. 5, 171).

Nolte, G., A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller (2008). "Robustly Estimating the Flow Direction of Information in Complex Physical Systems". In: *Physical Review Letters* 100.23, pp. 1–4 (cit. on p. 18).

Palmén, E. and C. W. Newton (1969). *Atmospheric circulation systems: Their structure and physical interpretation*. 13th ed. New York: Academic Press (cit. on pp. 127, 132).

Paluš, M. (1996). "Coarse-grained entropy rates for characterization of complex time series". In: *Physica D: Nonlinear Phenomena* 93.1-2, pp. 64–77 (cit. on p. 60).

Paluš, M., D. Hartman, J. Hlinka, and M. Vejmelka (2011). "Discerning connectivity from dynamics in climate networks". In: *Nonlinear Processes in Geophysics* 18.5, pp. 751–763 (cit. on p. 13).

Paluš, M. and M. Vejmelka (2007). "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections". In: *Physical Review E* 75.5, p. 056211 (cit. on p. 19).

Paluš, M., V. Komárek, Z. Hrnčír, and K. Štěrbová (2001). "Synchronization as adjustment of information rates: Detection from bivariate time series". In: *Physical Review E* 63.4, p. 46211 (cit. on p. 19).

Pant, G. B. and K. R. Kumar (1997). *Climates of South Asia*. New York: John Wiley and Sons Inc. (cit. on pp. 2, 28, 151).

Parthasarathy, B., A. Munot, and D. Kothawale (1994). "All-India monthly and seasonal rainfall series: 1871â1993". In: *Theoretical and Applied Climatology* 49.4, pp. 217–224 (cit. on p. 152).

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press (cit. on pp. 4, 16, 17, 19, 27, 31, 35, 103).

– (2009). "Causal inference in statistics: An overview". In: *Statistics Surveys* 3, pp. 96–146 (cit. on pp. 16, 31).

Permuter, H. H. and I. Naiss (2010). "Extension of the Blahut-Arimoto algorithm for maximizing directed information". In: *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1442–1449 (cit. on p. 113).

Peters, J., D. Janzing, and B. Schölkopf (2013). "Causal Inference on Time Series using Restricted Structural Equation Models". In: *Advances in Neural Information Processing Systems*, pp. 154–162 (cit. on pp. 17, 24).

Philander, S. G. H. (1990). *El Niño, La Niña, and the southern oscillation*. San Diego: Academic press (cit. on pp. 2, 125).

Pikovsky, A., M. Rosenblum, and J. Kurths (2003). *Synchronization: a universal concept in nonlinear sciences*. Vol. 12. Cambridge: Cambridge Univ Press (cit. on pp. 12, 18).

Póczos, B. and J. Schneider (2012). *Conditional Distance Variance and Correlation*. Tech. rep. Pittsburgh: Carnegie Mellon University (cit. on pp. 35, 64).

Póczos, B., Z. Ghahramani, and J. Schneider (2012). "Copula-based kernel dependency measures". In: *preprint arXiv:1206.4682* (cit. on pp. 33, 35, 64, 66).

Pompe, B. (1993). "Measuring statistical dependences in a time series". In: *Journal of Statistical Physics* 73.3, pp. 587–610 (cit. on p. 39).

– (1998). "Ranking and Entropy Estimation in Nonlinear Time Series Analysis". In: *Nonlinear analysis of physiological data*. Berlin Heidelberg: Springer. Chap. I, pp. 67–90 (cit. on p. 62).

– (2002). "Mutual information and relevant variables for predictions". In: *Modelling and Forecasting Financial Data*. Vol. 2. New York: Springer US, pp. 61–92 (cit. on pp. 157, 158, 161).

Pompe, B. and J. Runge (2011). "Momentary information transfer as a coupling measure of time series". In: *Physical Review E* 83.5, pp. 1–12 (cit. on pp. 19, 20, 36, 37, 51).

Prokopenko, M., J. Lizier, and D. Price (2013). "On Thermodynamic Interpretation of Transfer Entropy". In: *Entropy* 15.2, pp. 524–543 (cit. on pp. 114, 115).

Prusseit, J. and K. Lehnertz (2008). "Measuring interdependences in dissipative dynamical systems with estimated Fokker-Planck coefficients". In: *Physical Review E* 77.4, pp. 1–10 (cit. on p. 17).

Radebach, A., R. V. Donner, J. Runge, J. F. Donges, and J. Kurths (2013). "Disentangling different types of El Niño episodes by evolving climate network analysis". In: *Physical Review E* 88.5, p. 052807 (cit. on p. 13).

Ragwitz, M. and H. Kantz (2002). "Markov models from data by simple nonlinear time series predictors in delay embedding spaces". In: *Physical Review E* 65.5, p. 056201 (cit. on pp. 55, 157).

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan (2003). "Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century". In: *Journal of Geophysical Research* 108.D14, p. 4407 (cit. on pp. 152, 162, 205).

Rehfeld, K. and J. Kurths (2014). "Similarity estimators for irregular and age-uncertain time series". In: *Climate of the Past* 10.1, pp. 107–122 (cit. on p. 30).

Rehfeld, K., N. Marwan, J. Heitzig, and J. Kurths (2011). "Comparison of correlation analysis techniques for irregularly sampled time series". In: *Nonlinear Processes in Geophysics* 18.3, pp. 389–404 (cit. on p. 30).

Reichenbach, H. (1956). *The Direction of Time*. Berkeley and Los Angeles: University of California Press (cit. on pp. 2, 4).

Reid, N. (1995). "The roles of conditioning in inference". In: *Statistical Science* 10.2, pp. 137–230 (cit. on pp. 5, 56, 107, 171).

Rényi, A. (1959). "On measures of dependence". In: *Acta Mathematica Hungarica* 10.3-4, pp. 441–451 (cit. on pp. 2, 34).

– (1961). "On measures of entropy and information". In: *Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley: University of California Press, pp. 547–561 (cit. on p. 38).

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti (2011). "Detecting Novel Associations in Large Data Sets". In: *Science* 334.6062, pp. 1518–1524 (cit. on pp. 2, 33–35, 68, 71).

Riedl, M., A. Suhrbier, H. Stepan, J. Kurths, and N. Wessel (2010). "Short-term couplings of the cardiovascular system in pregnant women suffering from pre-eclampsia". In: *Philosophical Transactions of the Royal Society of London A* 368.1918, pp. 2237–50 (cit. on p. 17).

Risbey, J. S. and M. J. Pook (2009). "On the remote drivers of rainfall variability in Australia." In: *Monthly Weather Review* 137.10, pp. 3233–3253 (cit. on p. 150).

Robins, J. M., R. Scheines, P. Spirtes, and L. Wasserman (2003). "Uniform consistency in causal inference". In: *Biometrika* 90.3, pp. 491–515 (cit. on p. 30).

Rodionov, S. N. (2006). "Use of prewhitening in climate regime shift detection". In: *Geophysical Research Letters* 33.12, pp. 1–4 (cit. on p. 72).

Romano, M., M. Thiel, J. Kurths, and C. Grebogi (2007). "Estimation of the direction of the coupling by conditional probabilities of recurrence". In: *Physical Review E* 76.3, pp. 1–9 (cit. on p. 18).

Rosenblum, M. G., A. Pikovsky, and J. Kurths (1996). "Phase synchronization of chaotic oscillators". In: *Physical Review Letters* 76.11, p. 1804 (cit. on p. 18).

Rothman, P. (1999). *Nonlinear time series analysis of economic and financial data*. New York: Springer (cit. on p. 16).

Rowntree, P. R. (1972). "The influence of tropical east Pacific Ocean temperatures on the atmosphere". In: *Quarterly Journal of the Royal Meteorological Society* 98.416, pp. 290–321 (cit. on p. 136).

Rulkov, N. F., M. M. Sushchik, L. S. Tsimring, and H. D. I. Abarbanel (1995). "Generalized synchronization of chaos in directionally coupled chaotic systems". In: *Physical Review E* 51.2, p. 980 (cit. on p. 18).

Runge, J. (2013). "On the graph-theoretical interpretation of Pearson correlations in a multivariate process and a novel partial correlation measure". In: *preprint arXiv:1310.5169*, pp. 1–20 (cit. on pp. 171, 194, 205).

Runge, J., J. Heitzig, V. Petoukhov, and J. Kurths (2012a). "Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy". In: *Physical Review Letters* 108.25, pp. 1–4 (cit. on pp. 12, 20, 37, 46, 59, 126, 170, 171, 178).

Runge, J., J. Heitzig, N. Marwan, and J. Kurths (2012b). "Quantifying Causal Coupling Strength: A Lag-specific Measure For Multivariate Time Series Related To Transfer Entropy". In: *Physical Review E* 86.6, pp. 1–15 (cit. on pp. 37, 49, 51, 88, 93, 103, 113, 170, 173, 206).

Runge, J., V. Petoukhov, and J. Kurths (2014). "Quantifying the strength and delay of climatic interactions: the ambiguities of cross correlation and a novel measure based on graphical models". In: *Journal of Climate* 27.2, pp. 720–739 (cit. on pp. 12, 88, 113, 126, 128, 132, 134, 173, 204).

Runge, J. (2010). "Coupling in the Climate System". Diploma Thesis. Humboldt University (cit. on p. 19).

Russell, B (1912). "On the notion of cause". In: *Proceedings of the Aristotelian society*, pp. 1–26 (cit. on p. 31).

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata (1999). "A dipole mode in the tropical Indian Ocean". In: *Nature* 401.6751, pp. 360–363 (cit. on pp. 150, 153).

Schäfer, C., M. G. Rosenblum, J. Kurths, and H. H. Abel (1998). "Heartbeat synchronized with ventilation". In: *Nature* 392, pp. 239–240 (cit. on p. 18).

Scheffer, M., J. Bascompte, W. A. Brock, V. Brovkin, S. S. R. Carpenter, V. Dakos, H. Held, E. H. E. van Nes, M. Rietkerk, and G. Sugihara (2009). "Early-warning signals for critical transitions." In: *Nature* 461.7260, pp. 53–9 (cit. on pp. 153, 154).

Schelter, B., M. Winterhalder, R. Dahlhaus, J. Kurths, and J. Timmer (2006). "Partial phase synchronization for multivariate synchronizing systems". In: *Physical Review Letters* 96.20, p. 208103 (cit. on pp. 18, 22).

Schelter, B., J. Timmer, and M. Eichler (2009). "Assessing the strength of directed influences among neural signals using renormalized partial directed coherence." In: *Journal of neuroscience methods* 179.1, pp. 121–30 (cit. on p. 35).

Schiff, S. J., P. So, T. Chang, R. E. Burke, and T. Sauer (1996). "Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble". In: *Physical Review E* 54.6, p. 6708 (cit. on p. 18).

Schleussner, C. F., J. Runge, J. Lehmann, and A. Levermann (2014). "The role of the North Atlantic overturning and deep ocean for multi-decadal global-mean-temperature variability". In: *Earth System Dynamics* 5, pp. 103–115 (cit. on pp. 172, 173).

Schneider, J. and B. Póczos (2012). "Nonparametric estimation of conditional information and divergences". In: *15th International Conference on Artificial Intelligence and Statistics* XX, pp. 914–923 (cit. on p. 61).

Schreiber, T. (2000). "Measuring information transfer". In: *Physical Review Letters* 85.2, pp. 461–464 (cit. on pp. 19, 45, 46).

Schreiber, T. and H. Kantz (1995). "Noise in chaotic data: Diagnosis and treatment." In: *Chaos* 5.1, pp. 133–142 (cit. on p. 36).

Schwarz, G. (1978). "Estimating the dimension of a model". In: *The Annals of Statistics* 6.2, pp. 461–464 (cit. on p. 140).

Schweizer, B. and E. F. Wolff (1981). "On nonparametric measures of dependence for random variables". In: *The Annals of Statistics* 9.4, pp. 879–885 (cit. on pp. 34, 35).

Shannon, C. E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423 (cit. on pp. 4, 35, 38, 113).

Shannon, C. E. and W. Weaver (1963). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press (cit. on p. 35).

Shimizu, S. and P. O. Hoyer (2006). "A linear non-Gaussian acyclic model for causal discovery". In: *The Journal of Machine Learning Research* 7, pp. 2003–2030 (cit. on p. 17).

Simpson, S. L., F. D. Bowman, and P. J. Laurienti (2013). "Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain". In: *Statistics Surveys* 7, pp. 1–36 (cit. on p. 118).

Sims, C. A. (1972). "Money, income, and causality". In: *The American Economic Review* 62.4, pp. 540–552 (cit. on p. 16).

– (1980). "Comparison of interwar and postwar business cycles: Monetarism reconsidered". In: *The American Economic Review* 70.2 (cit. on p. 159).

Small, M and C. Tse (2002). "Minimum description length neural networks for time series prediction". In: *Physical Review E* 66.6, p. 066701 (cit. on p. 157).

Smirnov, D. A. (2013). "Spurious causalities with transfer entropy". In: *Physical Review E* 87.042917, pp. 1–12 (cit. on p. 83).

Smirnov, D. A. and B. Bezruchko (2009). "Detection of couplings in ensembles of stochastic oscillators". In: *Physical Review E* 79.4, p. 046204 (cit. on p. 18).

Smirnov, D. A. and I. Mokhov (2009). "From Granger causality to long-term causality: Application to climatic data". In: *Physical Review E* 80.1, p. 016208 (cit. on p. 17).

Solomon, S. (2007). *Climate change 2007–the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*. Cambridge: Cambridge University Press (cit. on pp. 2, 125).

Sommerlade, L., M. Eichler, M. Jachan, K. Henschel, J. Timmer, and B. Schelter (2009). "Estimating causal dependencies in networks of nonlinear stochastic dynamical systems". In: *Physical Review E* 80.5, p. 051128 (cit. on p. 17).

Soofi, E. S. (1994). "Capturing the intangible concept of information". In: *Journal of the American Statistical Association* 89.428 (cit. on p. 37).

Spirtes, P. and C. Glymour (1991). "An algorithm for fast recovery of sparse causal graphs". In: *Social Science Computer Review* 9.1, pp. 62–72 (cit. on pp. 4, 5, 23, 28).

Spirtes, P., T. Richardson, C. Meek, R. Scheines, and C. Glymour (1998). "Using path diagrams as a structural equation modeling tool". In: *Sociological methods & research* 27.2, p. 182 (cit. on p. 97).

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, prediction, and search*. Vol. 81. Boston: The MIT Press (cit. on pp. 4, 5, 19, 23, 24, 27–30, 103).

Staniek, M. and K. Lehnertz (2008). "Symbolic Transfer Entropy". In: *Physical Review Letters* 100.15, pp. 1–4 (cit. on p. 19).

Steinhaeuser, K., A. R. Ganguly, and N. V. Chawla (2012). "Multivariate and multiscale dependence in the global climate system revealed through complex networks". In: *Climate Dynamics* 39.3-4, pp. 889–895 (cit. on p. 12).

Steuer, R., J. Kurths, C. O. Daub, J. Weise, and J. Selbig (2002). "The mutual information: detecting and evaluating dependencies between variables". In: *Bioinformatics* 18 Suppl 2, S231–40 (cit. on p. 60).

Stögbauer, H., R. G. Andrzejak, A. Kraskov, and P. Grassberger (2004). "Reliability of ICA estimates with mutual information". In: *Independent Component Analysis and Blind Signal Separation*. Berlin Heidelberg: Springer, pp. 209–216 (cit. on p. 107).

Sugihara, G., R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch (2012). "Detecting causality in complex ecosystems." In: *Science* 338.6106, pp. 496–500 (cit. on pp. 18, 31).

Szegö, G. (1915). "Ein Grenzwertsatz über die Toeplitzschen Determinanten einer reellen positiven Funktion". In: *Mathematische Annalen* 76.4, pp. 490–503 (cit. on pp. 94, 183).

Szpiro, G. G. (1997). "Forecasting chaotic time series with genetic algorithms". In: *Physical Review E* 55.3, pp. 2557–2568 (cit. on p. 157).

Takens, F. (1981). "Detecting strange attractors in turbulence". In: *Dynamical systems and turbulence, Warwick 1980: Proceedings of a symposium held at the University of Warwick 1979-80*. Ed. by D. A. Rand and L.-S. Young. Vol. 898. Lecture Notes in Mathematics. Springer, New York, pp. 366–381 (cit. on pp. 18, 157).

Thompson, J. M. T. and J. Sieber (2011a). "Climate tipping as a noisy bifurcation: A predictive technique". In: *IMA Journal of Applied Mathematics* 76.1, pp. 27–46 (cit. on p. 153).

– (2011b). "Predicting climate tipping as a noisy bifurcation: A review". In: *International Journal of Bifurcation and Chaos* 21.2, pp. 399–423 (cit. on p. 153).

Thompson, J. M. T. and J. Sieber (2012). "Climate predictions: the influence of nonlinearity and randomness." In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 370.1962, pp. 1007–11 (cit. on p. 153).

Tillman, R. E. and P. Spirtes (2008). "When causality matters for prediction : investigating the practical tradeoffs". In: *Journal of Machine Learning Research Workshop and Conference Proceedings* 6, pp. 137–146 (cit. on p. 165).

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N. C. Lau, and C. Ropoleswki (1998). "Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures". In: *Journal of Geophysical Research: Oceans (1978-2012)* 103.C7, pp. 14291–14324 (cit. on p. 150).

Triacca, U. (2005). "Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature?" In: *Theoretical and Applied Climatology* 81, pp. 133–135 (cit. on p. 17).

Tsonis, A. A. and P. J. Roebber (2004). "The architecture of the climate network". In: *Physica A* 333, pp. 497–504 (cit. on pp. 11, 12, 173).

Tsonis, A. A. and K. L. Swanson (2008). "Topology and Predictability of El Nino and La Nina Networks". In: *Physical Review Letters* 100.22, p. 228502 (cit. on p. 157).

Tsonis, A. A., K. L. Swanson, and P. J. Roebber (2006). "What do networks have to do with climate?" In: *Bulletin of the American Meteorological Society* 87.5, pp. 585–595 (cit. on p. 12).

Tsonis, A. A., K. L. Swanson, and G. Wang (2008). "On the role of atmospheric teleconnections in climate". In: *Journal of Climate* 21.12, pp. 2990–3001 (cit. on pp. 2, 12, 118).

Tsujishita, T. (1995). "On triple mutual information". In: *Advances in Applied Mathematics* 16.3, pp. 269–274 (cit. on p. 42).

Van Rossum, G. and F. L. Drake Jr (1995). *Python reference manual*. Amsterdam: Centrum voor Wiskunde en Informatica (cit. on p. 7).

Vautard, R. and M. Ghil (1989). "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series". In: *Physica D* 35.3, pp. 395–424 (cit. on p. 12).

Vejmelka, M. and K. Hlaváčková-Schindler (2005). "Mutual Information Estimation in Higher Dimensions : A Speed-Up of a k -Nearest Neighbor Based Estimator". In: *Adaptive and Natural Computing Algorithms*. Ed. by B. Beliczynski, A. Dzielinski, M. Iwanowski, and B. Ribeiro. Lecture No. Vol. 226. 517133. Berlin Heidelberg: Springer (cit. on p. 62).

Vejmelka, M. and M. Paluš (2008). "Inferring the directionality of coupling with conditional mutual information". In: *Physical Review E* 77.2, p. 026214 (cit. on pp. 19, 61–63).

Verdes, P. (2005). "Assessing causality from multivariate time series". In: *Physical Review E* 72.2, p. 026222 (cit. on pp. 19, 29).

Vimont, D. J., J. M. Wallace, and D. S. Battisti (2002). "The Seasonal Footprinting Mechanism in the Pacific: Implications for ENSO". In: *Journal of Climate* 16.16, p2668 (cit. on p. 150).

Von Storch, H. and F. W. Zwiers (2002). *Statistical analysis in climate research*. Cambridge: Cambridge University Press (cit. on pp. 2, 12, 14, 15, 59, 64, 72, 87, 88, 107, 116, 139).

Walker, G. T. (1923). "Correlation in seasonal variations of weather, VIII: A Preliminary Study of World Weather". In: *Memoirs of the Indian Meteorological Department* 24.4, pp. 75–131 (cit. on pp. 4, 13, 136).

– (1924). "Correlation in Seasonal Variations of Weather, IX: A Further Study of World Weather". In: *Monthly Weather Review* 24.9, pp. 275–332 (cit. on pp. 13, 136).

Wallace, J. M. and D. S. Gutzler (1981). "Teleconnections in the 500 mb geopotential height field during the Northern Hemisphere winter". In: *Monthly Weather Review* 109, pp. 784–812 (cit. on p. 12).

Wang, C. (2002). "Atmospheric circulation cells associated with the El Niño-Southern Oscillation". In: *Journal of Climate* 15.4, pp. 399–419 (cit. on pp. 133, 136).

Wang, C., D. B. Enfield, S. Lee, and C. W. Landsea (2006). "Influences of the Atlantic warm pool on Western Hemisphere summer rainfall and Atlantic hurricanes". In: *Journal of Climate* 19.12, pp. 3011–3028 (cit. on p. 133).

Wang, Q., S. R. Kulkarni, and S. Verdú (2009). "Divergence Estimation for Multidimensional Densities Via k -Nearest-Neighbor Distances". In: *IEEE Transactions on Information Theory* 55.5, pp. 2392–2405 (cit. on p. 61).

Wang, X. L. (2008). "Accounting for Autocorrelation in Detecting Mean Shifts in Climate Data Series Using the Penalized Maximal t or F Test". In: *Journal of Applied Meteorology and Climatology* 47.9, pp. 2423–2444 (cit. on p. 72).

Webster, P. J. (1981). "Mechanisms determining the atmospheric response to sea surface temperature anomalies". In: *Journal of the Atmospheric Sciences* 38.3, pp. 554–571 (cit. on p. 136).

Webster, P. J. and S. Yang (1992). "Monsoon and ENSO: Selectively interactive systems". In: *Quarterly Journal of the Royal Meteorological Society* 118.507, pp. 877–926 (cit. on p. 165).

Webster, P. J. and C. D. Hoyos (2010). "Beyond the spring barrier?" In: *Nature Geoscience* 3.March, pp. 152 –153 (cit. on p. 165).

Wibral, M., N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente (2013). "Measuring information-transfer delays". In: *PloS one* 8.2, e55809 (cit. on pp. 20, 49, 55).

Wiener, N. (1956). *The theory of prediction*. New York: McGraw-Hill (cit. on p. 16).

Wright, S. (1921). "Correlation and causation". In: *Journal of Agricultural Research* 20.7, pp. 557–585 (cit. on pp. 2, 17).

– (1934). "The method of path coefficients". In: *The Annals of Mathematical Statistics* 5.3, pp. 161–215 (cit. on pp. 97, 101).

Wu, G., X. Duan, W. Liao, Q. Gao, and H. Chen (2011). "Kernel canonical-correlation Granger causality for multiple time series". In: *Physical Review E* 83.4, pp. 2–5 (cit. on p. 17).

Yakowitz, S. and M. Karlsson (1987). "Nearest neighbor methods for time series, with application to rainfall/runoff prediction". In: *Advances in the Statistical Sciences: Stochastic Hydrology*. Ed. by I. B. MacNeill, G. J. Umphrey, and A. I. McLeod. Springer Netherlands, pp. 149–160 (cit. on p. 157).

Yamasaki, K., A. Gozolchiani, and S. Havlin (2008). "Climate Networks around the Globe are Significantly Affected by El Nino". In: *Physical Review Letters* 100.22, p. 228501 (cit. on p. 13).

Zebiak, S. E. and M. A. Cane (1987). "A Model of El Niño-Southern Oscillation". In: *Monthly Weather Review* 115, pp. 2262–2278 (cit. on p. 157).

Zhang, K., J. Peters, D. Janzing, and B. Schölkopf (2012). "Kernel-based conditional independence test and application in causal discovery". In: *preprint arXiv:1202.3775* (cit. on pp. 17, 66).

Zhang, X. (2004). "Comment on'Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test' by Sheng Yue and Chun Yuan Wang". en. In: *Water Resources Research* 40.3, W03805 (cit. on p. 72).

Zhao, Y., S. Billings, H. Wei, and P. Sarrigiannis (2012). "Tracking time-varying causality and directionality of information flow using an error reduction ratio test with applications to electroencephalography data". In: *Physical Review E* 86.5, p. 051919 (cit. on p. 17).

Zou, Y., J. Heitzig, R. V. Donner, J. F. Donges, J. D. Farmer, R. Meucci, S. Euzzor, N. Marwan, and J. Kurths (2012). "Power-laws in recurrence networks from dynamical systems". In: *Europhysics Letters* 98.4, p. 48001 (cit. on p. 18).

Zwiers, F. W. (1990). "The Effect of Serial Correlation on Statistical Inferences Made with Resampling Procedures". In: *Journal of Climate* 3.12, pp. 1452–1461 (cit. on p. 72).